Statistical Methods for Clinical Trials with Multiple Outcomes, HIV Surveillance, and Nonparametric Meta-Analysis

Dissertation Advisor: Professor Lee-Jen Wei Brian Lee Claggett

Statistical methods for clinical trials with multiple outcomes, HIV surveillance, and nonparametric meta-analysis

Abstract

Central to the goals of public health are obtaining and interpreting timely and relevant information for the benefit of humanity. In this dissertation, we propose methods to monitor and assess the spread HIV in a more rapid manner, as well as to improve decisions regarding patient treatment options.

In Chapter 1, we propose a method, extending the previously proposed dual-testing algorithm and augmented cross-sectional design, for estimating the HIV incidence rate in a particular community. Compared to existing methods, our proposed estimator allows for shorter follow-up time and does not require estimation of the mean window period, a crucial, but often unknown, parameter. The estimator performs well in a wide range of simulation settings. We discuss when this estimator would be expected to perform well and offer design considerations for the implementation of such a study.

Chapters 2 and 3 are concerned with obtaining a more complete understanding of the impact of treatment in randomized clinical trials in which multiple patient outcomes are recorded. Chapter 2 provides an illustration of methods that may be used to address concerns of both risk-benefit analysis and personalized medicine simultaneously, with a goal of successfully identifying patients who will be ideal candidates for future treatment. Risk-benefit analysis is intended to address the multivariate nature of patient outcomes, while "personalized medicine" is concerned with patient heterogeneity, both of which complicate the determination of a treatment's usefulness. A third complicating factor is the duration of treatment use. Chapter 3 features proposed methods for assessing the impact of treatment as a function of time, as well as methods for summarizing the impact of treatment across a

range of follow-up times.

Chapter 4 addresses the issue of meta-analysis, a commonly used tool for combining information for multiple independent studies, primarily for the purpose of answering a clinical question not suitably addressed by any one single study. This approach has proven highly useful and attractive in recent years, but often relies on parametric assumptions that cannot be verified. We propose a non-parametric approach to meta-analysis, valid in a wider range of scenarios, minimizing concerns over compromised validity.

Contents

Acknowledgments

I would like to thank each of the members of the Department of Biostatistics who served as a mentor and/or advisor during my time as a student. Both of my primary advisors, Steve Lagakos and LJ Wei, helped to provide an excellent graduate experience, supplying ample time, advice, and enthusiasm for research. Special thanks goes to Rui Wang, who led me to work with both Steve and LJ, and who exceeded expectations by stepping in to help ensure that the work begun by Steve was seen to completion, and for providing personal and professional guidance. I would further like to thank my final committee member Tianxi Cai, who served as my primary source for filling in any and all gaps in my statistical knowledge.

Thanks to Jelena Follweiler, without whom I probably would not be graduating, for making sure all those forms got filled out and turned in on time. Further thanks to my professors, classmates, co-authors, and parents.

Finally, I thank my fiancee Mallory for keeping me sane during the entire dissertation process, for finding a tie for me to wear with less than one hour to go before my defense, and particularly for suggesting, about six years ago, that I figure out what "biostatistics" means and that I consider applying to graduate programs.

Augmented Cross-Sectional Studies with Abbreviated Follow-up for Estimating HIV Incidence

Brian Claggett, Stephen W. Lagakos, and Rui Wang

Department of Biostatistics
Harvard School of Public Health

1.1 Introduction

Cross-sectional HIV incidence estimation based on a sensitive and less-sensitive test offers great advantages over the traditional cohort study for incidence estimation (Brookmeyer and Quinn, 1995; Janssen et al., 1998). Based on features of the evolving HIV-1 antibody response during the months after primary HIV-1 infection, several enzyme-immunoassays (EIAs) have been developed to differentiate early and long-standing HIV infections. Subjects who test positive on an EIA sensitive assay but negative on a less-sensitive assay are identified as early infections while subjects who are reactive on both are considered to be long-term HIV infections. An incidence estimate can be obtained by equating incidence with prevalence of persons with early infection divided by the time between seroconversion on the 2 tests (commonly termed "window period"). This approach has attracted much attention. One popular assay is the IgG capture BED EIA (BED) assay (Parekh et al., 2002). Balasubramanian and Lagakos (2010) postulate a longitudinal history statistical model of HIV seconversion and subsequent reactivity to a less sensitive assay, and show that the cross-sectional estimator arises as a maximum likelihood estimator under this model. They further extend this method to allow the assessment of covariate effects. However, concerns have been raised in recent years regarding the reliability of this method (Sakarovitch et al., 2007). Some less-sensitive tests, for example, the BED assay, may yield false negative results; that is, subjects may remain negative to the less-sensitive assay long after they have been infected (Hargrove et al., 2008). It has been noted that if such a subpopulation exists, then the original cross-sectional incidence estimator must be adjusted to account for this fact (Brookmeyer, 2009; Welte et al., 2009; Wang and Lagakos, 2009).

Recently, Wang and Lagakos (2010) proposed an augmented cross-sectional study design which adds a longitudinal component to the traditional cross-sectional study by following forward those subjects who test positive on the sensitive assay and test negative on the less-sensitive assay until either they become reactive to the less-sensitive assay or for a predetermined length of follow-up time τ, whichever occurs first. The augmented cross-

sectional study provides one way to estimate the size of the subpopulation who will remain negative permanently on the less-sensitive test. An estimate of the false negative rate can be obtained by calculating the proportion of subjects who remain negative at the end of the follow-up period. An underlying assumption for an unbiased estimate of the false negative rate is that τ is chosen to be long enough so that those who have not become reactive to the less-sensitive assay by the end of the follow-up period will remain negative indefinitely. Thus, the choice of τ requires knowledge of an upper limit for the window period (W^*) among those who will become reactive. Two issues arise. First, does there exist a subpopulation where infected subjects would remain nonreactive to the less-sensitive assay permanently? If so, what is the upper limit of the window period among those who will become reactive? Although there is some belief that those who will become reactive to the BED assay will do so within a year, considerable uncertainty regarding this upper limit exists.

In this paper, we assess the impact of varying follow-up time on estimating HIV incidence within the context of an augmented cross-sectional design, evaluate the robustness of incidence estimators to the assumptions about the existence and size of the subpopulation where infected subjects will remain negative permanently on the less-sensitive test, and propose a new estimator based on abbreviated follow-up time (AF). In Section 2.1, we generalize the likelihood function in Wang and Lagakos (2010) by incorporating length of follow-up and analyze the behavior of the estimators when different underlying models are assumed. In Section 2.2, we introduce the AF estimator, which allows shorter follow-up time than the original estimator based on the augmented design and does not require estimation of the window period. In Section 3, we use simulations to compare the performance of various estimators. In Section 4, we discuss areas of future research. All proofs are relegated to Appendices.

1.2 Methods

1.2.1 Notation and Model

We consider the four-state progressive-disease model as in Section 3 of Wang and Lagakos (2010). Here, State 1 represents the pre-seroconversion state (uninfected or infected but not detectable by the sensitive test). State 2 represents the "recent infection" state, in which an infected subject is detectable by the sensitive diagnostic test, but not yet by the less-sensitive test (though he/she will eventually test positive). State 3 represents the "nonrecent infection" state in which an infected individual is detectable by both the sensitive and less-sensitive diagnostic tests. State 4 represents subjects who will remain nonreactive to the less-sensitive assay permanently.

Using the same notation as in Wang and Lagakos (2010), let time 0 denote birth of an individual and T denote the calendar time of HIV seroconversion. Let $f(u)$, $\lambda(u)$, and $F(u)$ denote the density, incidence rate, and cumulative distribution functions of T at time $u \geq 0$. Let W denote the individual's time spent in State 2, with cumulative distribution function $G(\cdot)$. Denote the upper limit of support for W by W^*. Let $\mu = E(W)$ denote the mean time from infection to testing positive on less-sensitive test, amongst the population who will test positive eventually. Let p denote the proportion of such patients among all infected subjects. Let t denote the calendar time of the cross-sectional sample. As in Wang and Lagakos (2010), we assume that T is independent of W and $f(u) = f$, for $u \in (t - W^*, t)$. Our goal is to make inferences about $\lambda(t) = f/\{1 - F(t)\}$, the HIV incidence rate at the time, t, of the cross-sectional sample. Hereafter we denote $\lambda(t)$ as λ for simplicity because the remainder of the paper focuses on a fixed calendar time t, the time at which the cross-sectional sample is obtained.

Suppose that a random sample of size N is drawn from a population of asymptomatic individuals at calendar time t, and tested using both a sensitive and less-sensitive diagnostic

test. Let N_1, N_2, and N_3 denote the numbers of subjects who test negative on both tests (State 1), positive only on the sensitive test (State 2 or 4), and positive on both tests (State 3), respectively. In an augmented cross-sectional study, we will follow and periodically retest the N_2 subjects who test negative on the less-sensitive test for some time period τ, at which time we will have observed $n_{1\tau}$ transitions to the "non-recent infection" state. Let $n_{0\tau} = N_2 - n_{1\tau}$ denote the number of remaining subjects, including subjects who may become positive on the less-sensitive test at a later time as well as those who will remain negative permanently. Let X denote the forward recurrence time, which represents the time elapsed from t until entrance into State 3. Let $h(x \mid t)$ and $H(x \mid t)$ denote the conditional density and cumulative function of X. Each of the $n_{1\tau}$ subjects gives rise to an interval censored observation $[a_i, b_i]$ of the forward recurrence time X_i, where a_i denotes the elapsed time between t and the last negative test result and b_i denotes the elapsed time between t and the first positive test result. Let n_1 and n_0 denote the true number of subjects who are in State 2 and 4, respectively, at time t.

Let $\pi_1(t)$, $\pi_2(t)$, $\pi_3(t)$, and $\pi_4(t)$ denote the prevalence probabilities in State 1, 2, 3, and 4 at time t, respectively. The likelihood function corresponding to an arbitrary τ is given by (see Appendix 1):

$$\begin{aligned} L =& \{\pi_1(t)\}^{N_1}\{\pi_2(t)H(\tau \mid t)\}^{n_{1\tau}}\{\pi_3(t)\}^{N_3}[\pi_4(t) + \pi_2(t)\{1 - H(\tau \mid t)\}]^{n_{0\tau}} \\ & \times \prod_{i=1}^{n_{1\tau}} \frac{H(b_i \mid t) - H(a_i \mid t)}{H(\tau \mid t)} \\ =& \phi^{N_1}\{p(1-\phi-\phi\lambda\mu)\}^{N_3}(\phi\lambda p\mu)^{n_{1\tau}}[(1-p)(1-\phi) + \phi\lambda\mu p\{1 - H(\tau \mid t)\}]^{n_{0\tau}} \\ & \times \prod_{i=1}^{n_{1\tau}} \{H(b_i \mid t) - H(a_i \mid t)\} \end{aligned} \tag{1.1}$$

where $\phi = 1 - F(t)$ and

$$H(x \mid t) = \int_0^x \frac{1 - G(v)}{\mu} dv \tag{1.2}$$

as shown in Wang and Lagakos (2010). This relationship implies that $H(\cdot)$ is uniquely determined once $G(\cdot)$ is specified. Because the right hand side of (1.2) is free of t, and the

remainder of the paper focuses on a fixed time t, we hereafter use $H(x)$ to represent $H(x \mid t)$ for simplicity. We note that without parametric assumptions about the distribution $G(\cdot)$, the parameters (μ, p) are non-identifiable, as the true state of subjects who have yet to transition to State 3 by the end of follow-up cannot be known. Even with a parametric assumption, the parameters (μ, p) are weakly identifiable, in the sense that the likelihood function is relatively flat near the true parameter values, resulting in a near-singular information matrix. Below we attempt to address this issue by making additional assumptions on either p or $G(\tau)$.

Assume $p = 1$

If all infected subjects will eventually test positive on the less-sensitive test, then $p = 1$ and the above likelihood reduces to

$$\begin{aligned} L =& \phi^{N_1}(1-\phi-\phi\lambda\mu)^{N_3}(\phi\lambda\mu)^{n_{1\tau}} \prod_{i=1}^{n_{1\tau}} \{H(b_i) - H(a_i)\}[\phi\lambda\mu\{1-H(\tau)\}]^{n_{0\tau}} \\ =& \phi^{N_1}(1-\phi-\phi\lambda\mu)^{N_3}(\phi\lambda\mu)^{N_2} \prod_{i=1}^{N_2} \{H(b_i) - H(a_i)\}. \end{aligned} \tag{1.3}$$

For those subjects who have not become reactive to the less-sensitive assay at time τ, we let $a_i = \tau$ and $b_i = \infty$. The above likelihood is given in equation (5) of Wang and Lagakos (2010). The assumption of $p = 1$ implies that all recently infected individuals will eventually test positive on the less-sensitive test. If this assumption is true, then the resulting estimator

$$\hat{\lambda}_A = \frac{N_2}{N_1 \hat{\mu}} \tag{1.4}$$

(hereafter referred to as Estimator A) is consistent regardless of follow-up time τ, provided that the parametric assumption about the underlying distribution for sojourn time is correct. Here we use $\hat{\mu}$ to denote the maximum likelihood estimate of μ.

However, if this assumption of $p = 1$ is violated, then it was shown in Wang and Lagakos (2009) that this estimator $\hat{\lambda}_A$ converges in probability to

$$\frac{p\phi\lambda\mu + (1-p)(1-\phi)}{\phi\mu_{\tau A}} \tag{1.5}$$

where $\mu_{\tau A}$ is the probability limit of the (biased) estimate of μ resulting from τ units of follow-up and the (improper) assumption that $p = 1$.

Under mild violations of this assumption ($p = .99$), we find that the asymptotic bias is modest for $\tau \leq \mu$, but then increases quickly, with $E(\hat{\lambda}_A)/\lambda > 1.25$ when $\tau \geq 2\mu$ under many plausible scenarios, with the degree of bias increasing with the population prevalence, leading to a significant overestimation of the true incidence rate.

Assume that follow-up time is longer than W^*

When the follow-up time τ is longer than W^*, we have $H(\tau) = 1$, and the likelihood function (1.1) reduces to

$$L = \phi^{N_1}\{p(1-\phi-\phi\lambda\mu)\}^{N_3}(\phi\lambda p\mu)^{n_{1\tau}}\{(1-p)(1-\phi)\}^{n_{0\tau}}\prod_{i=1}^{n_{1\tau}}\{H(b_i)-H(a_i)\}. \tag{1.6}$$

This corresponds to the scenario described in Section 3 of Wang and Lagakos (2010) when all individuals in State 2 are followed until they have entered State 3. If we know W^* up to an upper limit, then we can choose $\tau > W^*$, so that $H(\tau) = 1$. The resulting estimator,

$$\hat{\lambda}_B = \frac{n_{1\tau}}{N_1\hat{p}_\tau\hat{\mu}_\tau} \tag{1.7}$$

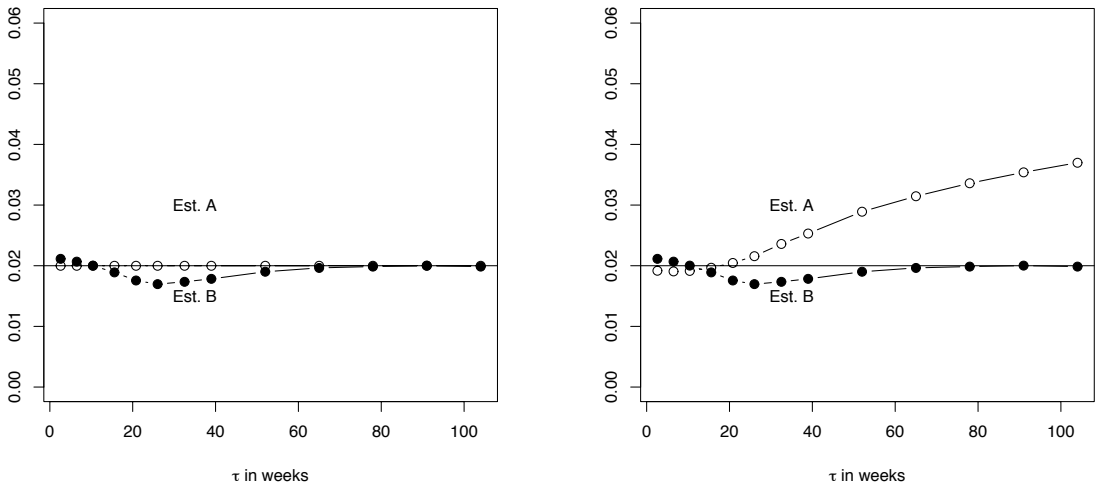
(hereafter referred to as Estimator B) is consistent, where $\hat{p}_\tau = (N_3 + n_{1\tau})/(N_3 + N_2)$.

If this assumption $\tau > W^*$ is violated, then some of the $n_1 - n_{1\tau}$ truly recently infected individuals who have not transitioned during the follow-up period may be incorrectly classified as "long-term non-progressors". It is shown in Appendix 2 that this estimator converges in probability to

$$\lambda\frac{\mu}{\mu_{\tau B}}\frac{(1-\phi)H(\tau)}{1-\phi-\phi\lambda\mu\{1-H(\tau)\}} \tag{1.8}$$

where $\mu_{\tau B}$ is the probability limit of the (biased) estimate of μ resulting from τ units of follow-up and the (improper) assumption that $H(\tau) = 1$.

In Figure 1.1, we show an example of the probability limits of Estimators A and B as a function of follow-up time τ, for $p = 1$ (Fig. 1.1(a)) and $p = .95$ (Fig. 1.1(b)). For this example, we assumed prevalence and incidence values of 15% and 2%, respectively, with mean window period of 6 months and the sojourn times following a Weibull distribution with shape parameter $k = 2$. We used (1.5) and (1.8) to calculate the probability limits at each time τ. Asymptotic values of $\hat{\mu}_{\tau A}$ and $\hat{\mu}_{\tau B}$ were estimated by simulating data representing approximately one million subjects with biweekly follow-up. The asymptotic bias of Estimator A is seen to be quite sensitive to departures from the assumption of $p = 1$, and increases as τ increases. The limit of Estimator A rises to 2.9% for $\tau = 52$ and 3.4% when $\tau = 78$ weeks. The asymptotic bias of Estimator B is comparatively small even when τ is much smaller than W^*.



(a) p=1

(b) p=0.95

Figure 1.1: *Probability Limits of Estimator A (open circles) and Estimator B (solid circles) for Various τ and p, with $\phi = .85$, $\lambda = .02$, and Assuming that $G(\cdot)$ is Weibull with Mean .5 and Shape Parameter $k = 2$.*

Assume $p = p_0 < 1$

If the assumption of $p = 1$ is likely to be violated, and we are unable to posit a value for W^*, we can assume a pre-specified value of p (denoted p_0), possibly on the basis of prior studies in a similar population, in which case the likelihood becomes

$$\begin{aligned} L =& \phi^{N_1}\{p_0(1-\phi-\phi\lambda\mu)\}^{N_3}[(1-p_0)(1-\phi)+\phi\lambda\mu p_0\{1-H(\tau)\}]^{n_{0\tau}}(\phi\lambda p_0\mu)^{n_{1\tau}} \\ & \times \prod_{i=1}^{n_{1\tau}}\{H(b_i)-H(a_i)\} \end{aligned} \tag{1.9}$$

To obtain the MLEs for (λ, ϕ, μ), we can again assume some parametric distribution for $G(\cdot)$ and numerically maximize L. We refer to this maximum likelihood estimate of λ, $\hat{\lambda}_C$, as Estimator C. An estimate of the covariance matrix of (λ, ϕ, μ) is provided by the sample Fisher information corresponding to L in (1.9). In Web Appendix A, we derive the estimated covariance matrix assuming that $G(\cdot)$ follows a Weibull distribution, a flexible distribution with support on positive numbers which is commonly-used in parametric survival modeling. This estimator is consistent when p_0 and the parametric assumption about $G(\cdot)$ are correctly specified, regardless of length of follow-up τ. In Section 3, we examine the robustness of this estimator to violation of assumptions about p_0 and parametric distributions through simulations. For estimators A, B, and C, a $(1-\alpha)100\%$ confidence interval is given by

$$(\hat{\lambda}e^{-Z_{1-\alpha/2}\hat{s}/\hat{\lambda}}, \hat{\lambda}e^{Z_{1-\alpha/2}\hat{s}/\hat{\lambda}}) \tag{1.10}$$

where $\hat{s}$ is the square root of the diagonal element of the sample Fisher information corresponding to $\hat{\lambda}$. The confidence intervals in (1.10) are additive on the log scale. As suggested by the associate editor, we also consider the use of profile likelihood confidence intervals in Section 3.

1.2.2 AF Estimator

Motivation

If we were to conduct a cohort study to estimate HIV incidence, i.e., the rate of transition from State 1 to State 2 in our disease model, we could estimate the incidence rate by # of events/person-years, where "number of events" refers to the number of transitions from State 1 to State 2, and "person-years" refers to the cumulative time contributed by the uninfected individuals in the cohort. Let T_2 denote the time of entering State 3 and $f_2(\cdot)$ denotes its corresponding density function. Under the assumption of a constant infection density, $f(u) = f$ for $u \in (t - W^*, t)$, the probability of transition from State 1 to State 2 at time t, $f(t)$, is related to the probability of transition from State 2 to State 3, $f_2(t)$. In Appendix 3 we show that $f(t) = f_2(t)/p$, which implies that

$$\lambda = \frac{f}{1 - F(t)} = \frac{f_2(t)}{p\{1 - F(t)\}}. \tag{1.11}$$

Assume $f_2(u) = f_2$, for $u \in (t, t + \tau)$. This is reasonable for a relatively small τ (e.g., 12 weeks). From (1.11), we have $\lambda = f_2/[p\{1 - F(t)\}]$. Suppose there is some minimum time, say, τ^*, spent in the recent infection state. That is, an infected individual will test negative on the less-sensitive assay for at least τ^* from seroconversion. If the follow-up time τ is chosen such that τ is less than this minimum time τ^*, then $G(\tau) = 0$. This implies that P(Subject in State 1 at t | Subject in State 3 at $t + \tau$) $= 0$. Hence, if we are interested in observing the number of transitions into State 3 that occur in our sample between the time of the cross-sectional testing, t, and the end of follow-up, $t + \tau$, we only need to follow the N_2 patients who test positive on the sensitive test and negative on the less-sensitive test at time t. The total number of transitions into State 3 during interval $(t, t + \tau)$ is given by $n_{1\tau}$ and f_2 can be estimated by $n_{1\tau}/(N\tau)$. This, combined with the fact that $F(t)$ can be estimated by $(N - N_1)/N$ Wang and Lagakos (2009), motivates us to consider the following

estimator:

$$\hat{\lambda} = \frac{n_{1\tau}}{N_1 \hat{p} \tau}, \tag{1.12}$$

where

$$\hat{p} = \frac{N_3 + n_{1\tau}}{N_3 + N_2}, \tag{1.13}$$

as used in Estimator B. We note that $\hat{p}$ will usually underestimate the true value of p because the use of a shorter follow-up time means that $n_{1\tau}$ will underestimate the true number of transitions, n_1, to State 3, had we been able to follow all N_2 subjects found to be in the recent state for W^*. However, this will help to compensate for the fact that $G(\tau)$ may not be strictly equal to 0 and hence $n_{1\tau}$ could also underestimate the true number of transitions to State 3 had we been able to follow all $N_1 + N_2$ subjects "at risk" for transition into State 3 during the interval $(t, t + \tau)$. We refer to this design as an augmented cross-sectional design with abbreviated follow-up and the corresponding estimator as the AF estimator.

We note that this estimator has the same form as Estimator B, with the value $\hat{\mu} = \tau$. In fact, if τ is small, but we continue to use observed forward recurrence times for the $n_{1\tau}$ subjects to estimate the mean and the shape parameter (k) of a Weibull distribution $G(\cdot)$ as described in Estimator B, we find through simulations that $\hat{k}$ tends to be quite large and $\hat{\mu} \approx \tau$.

Properties

In Appendix 4 we show that

$$E(\hat{\lambda}_{AF}) \approx \frac{\lambda \mu H(\tau)}{\tau} \frac{1}{1 - p_r(1 - H(\tau))}, \tag{1.14}$$

where

$$p_r = \frac{\lambda \phi \mu}{1 - \phi} = P(\text{recent infection} \mid \text{currently infected}). \tag{1.15}$$

In a region with high prevalence and incidence, for example, $(\lambda, \phi, \mu) = (.04,\ .7,\ .5)$, we have $p_r < .05$. As another example, a region with low prevalence and incidence with $(\lambda, \phi, \mu) = (.01, .92, .5)$ would yield $p_r < .06$. Because p_r is small and for small τ, $H(\tau) \approx \tau h(0) = \tau/\mu$, we have

$$E(\hat{\lambda}_{AF}) \approx \frac{\lambda\mu H(\tau)}{\tau} \frac{1}{1 - p_r\{1 - H(\tau)\}} \approx \frac{\lambda\mu\tau h(0)}{\tau} = \frac{\lambda\mu\tau}{\tau\mu} = \lambda. \tag{1.16}$$

It is shown in Appendix 5 that

$$\widehat{Var}\{log(\hat{\lambda}_{AF})\} \approx \frac{1}{n_{1\tau}}. \tag{1.17}$$

Thus, for early stopping times τ, a simple estimator and 95% CI could be calculated by counting the number of transitions occurring during the τ follow-up time, with the point estimate $\hat{\lambda}_{AF}$ given in (1.12) and confidence interval $(\hat{\lambda}_{AF}e^{-1.96\sqrt{\frac{1}{n_{1\tau}}}}, \hat{\lambda}_{AF}e^{1.96\sqrt{\frac{1}{n_{1\tau}}}})$.

For the AF estimator to be expected to be close to the true incidence λ by a factor of B, where $B < (1 - p_r)$, i.e. $B\lambda \leq E(\hat{\lambda}_{AF}) \leq \lambda/B$, we can choose τ such that (see Appendix 6)

$$\tau \leq \frac{\int_0^\tau 1 - G(x)\ dx}{B[1 - p_r\{1 - H(\tau)\}]}. \tag{1.18}$$

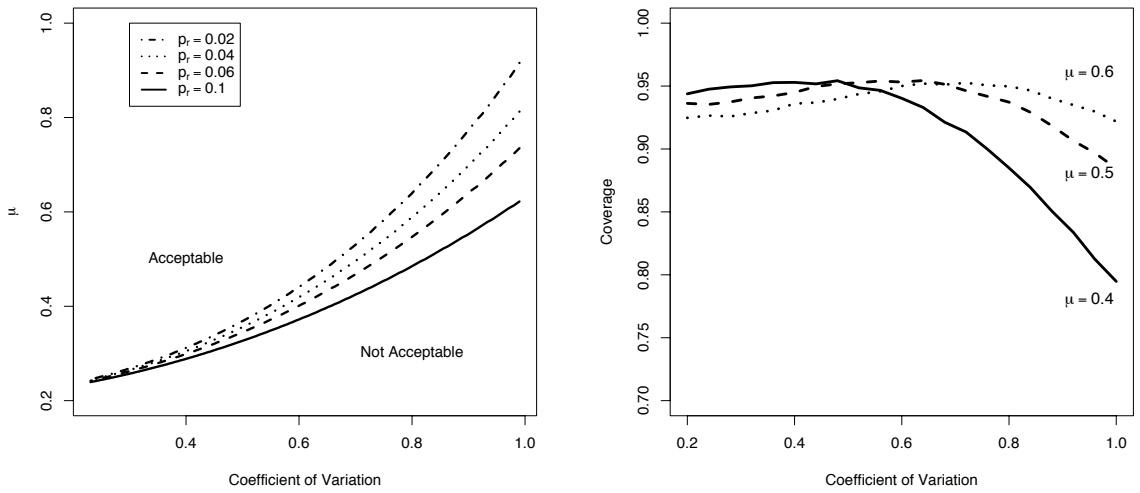
Because $[1 - p_r\{1 - H(\tau)\}] \leq 1$ and close to 1, a slightly more strict condition that will guarantee the boundedness of the bias of the AF Estimator is

$$\tau \leq \frac{\int_0^\tau 1 - G(x)\ dx}{B}. \tag{1.19}$$

The above inequality highlights the dependence of the AF estimator on the assumption of a relatively "flat" survival function $1 - G(x) : x \in (0, \tau)$, i.e. some minimum amount of time spent in State 2. A "perfect" test ($W = \mu$, with probability 1) would result in $g(x) = 0$ on the range (0,τ), and thus $\int_0^\tau 1 - G(x)\ dx \equiv \tau$ for any $\tau < \mu$. In practice, the less-sensitive test does not have this "perfect" property. We find that the bias of the AF estimator generally increases as the choice of stopping time τ increases. Because the width

of the confidence interval based on the AF estimator is purely "event-driven", it becomes smaller with each additional observed transition. Therefore, a larger value of τ is associated with increasing bias and decreasing variance. We have found through extensive simulations that $\tau \approx \mu/2$ is generally a good choice. Unless prior information is known about the value of μ in the population of interest, we recommend the choice of $\tau = 12$ weeks for use with the AF estimator. We show in Section 3 that this leads to good performance in a variety of settings.

In practice we do not know the exact distribution of sojourn times in a given population. Rather, we may have some idea of mean window period μ and possibly coefficient of variation $c = \sigma/\mu$. With a parametric assumption for the window period distribution and (μ, c), we can calculate $H(\tau)$ and verify whether a 12-week AF estimator satisfies $B\lambda \leq E(\hat{\lambda}_{AF}) \leq \lambda/B$ for a fixed B and a hypothetical value for p_r, which is determined by (λ, ϕ, μ) as in (1.15). In Figure 1.2(a), we provide the acceptable region, in terms of μ and c, of the 12-week AF estimator for $B = 0.9$ and $p_r = .02, .04, .06$, and .1. The boundaries presented in this plot are generated by finding $\mu_{min;c}$ that satisfies (1.18) for each fixed c. We note that for $\mu_1 < \mu_2$, if (μ_1,c) is in the acceptable region, then (μ_2, c) must also be in the acceptable region. Therefore, $\mu_{min;c}$ provides a lower bound for all μ that satisfies (1.18) for each c. We also find that the region which guarantees $B\lambda \leq E(\hat{\lambda}_{AF}) \leq \lambda/B$ when the true distribution is Weibull is somewhat smaller than the corresponding region when the true distribution is Lognormal. Because of this, we recommend using the reference region based on the Weibull assumption when deciding whether to use the AF estimator, in order to be conservative. An even more conservative boundary can be generated using (1.19), which does not dependent on p_r and hence does not require assumptions about (ϕ, λ). Figure 1.2(b) shows the actual coverage levels of the nominal 95% confidence intervals of the 12-week AF estimator for different values of (c, μ), in a setting where $(N, \lambda, \phi, p) = (6000, .03, .85, .99)$. We find that the actual coverage levels are close to the nominal level for all parameter pairs that fall within the acceptable region corresponding to B=.9, as shown in 1.2(a).



(a) Acceptable Region (b) Coverage

Figure 1.2: *Bias and Coverage of 12-week AF Estimator*
Left: *The 12-week AF Estimator will have acceptably small bias when (c,μ) lie above the line associated with desired bound B=0.90*
Right: *Actual Coverage of Nominal 95% CI,*
Coverage values obtained via simulation using $(\phi, \lambda, p) = (0.85, 0.03, 0.99)$ *and* $G(\cdot) \sim$ *Weibull*

Design Considerations

Suppose that one wanted to conduct a study to estimate the incidence of HIV in a region with some hypothesized parameter values (ϕ, λ, p), and had decided that the usage of the AF estimator is both desirable and appropriate. Under this proposed design plan, biweekly follow-up is not necessary, as only the total number of transitions occurring during the 12 weeks of follow-up is needed to produce this estimate of incidence, and could be obtained via a single follow-up visit 12 weeks (τ=12/52 years) after the initial testing. This designed study would require the administration of N (typically thousands) sensitive tests and $N - N_1$ less-sensitive tests at time t, and a further N_2 (typically < 100) less sensitive tests at time $t + \tau$. The precision of the AF estimator is closely tied with the number of transitions into State 3 within 12 weeks ($n_{1\tau}$). Below we show how to calculate the total sample needed so that the ratio of the true incidence rate and the estimated rate, $e^{Z_{\alpha/2}\sqrt{1/n_{1\tau}}}$, is smaller

than δ with $(1-\alpha)100\%$ confidence. To achieve this, we require that $n_{1\tau} \geq \{Z_{\alpha/2}/log(\delta)\}^2$. For example, for $\delta = 2$ and $\alpha = .05$, we would need the number of transitions $n_{1\tau}$ to be at least 8. If we would like to achieve better precision, and let $\delta = 1.5$, then we would require $n_{1\tau} \geq 24$.

Earlier we showed that the probability that a recently infected subject will transition to State 3 within τ units of follow-up is $\phi\lambda p\mu H(\tau) \approx \phi\lambda p\tau$ for small τ. The sample size N has the negative binomial distribution with number of failures $r = n_{1\tau}$ and success rate $P_s = 1 - \phi\lambda p\tau$. Therefore, the sample size N required to obtain the observation of at least $n_{1\tau}$ failures with probability β is $Q_{NB}(\beta, r = n_{1\tau}, P_s = 1 - \phi\lambda p\tau)$, where Q refers to the quantile function of negative binomial distribution. A close approximation to this value can be obtained by the following formula

$$\tilde{N} = \frac{(n_{1\tau} + Z_\beta \cdot \sqrt{n_{1\tau}})(1 - \phi\lambda p\tau)}{\phi\lambda p\tau}. \tag{1.20}$$

For example, in a high prevalence ($\phi = .75$) and high incidence ($\lambda = .04$) population, if we assume that $p = .94$, and would like to obtain at least 8 transitions into State 3 within 12 weeks with probability .9, then we would need to include at least $N = 1799$ subjects in the cross-sectional sample. Suppose, instead, that we are in a setting with lower prevalence ($\phi = .9$) and incidence ($\lambda = .02$) and wish to observe at least 24 transitions with probability .7. In this case, we would need $N = 6713$. Using the formula in (1.20) yields $\tilde{N}$=1775 and 6778, respectively, for the two settings discussed.

For a fixed sample size N and given parametric distribution $G(\cdot)$ and parameter values (ϕ, λ, p), we can find the optimal follow-up time τ that minimizes the mean squared error (MSE) of the AF Estimator numerically. For example, for the scenario where $\phi = .75$, $\lambda = .04$, and $N = 1799$, assuming that the true sojourn time distribution $G(\cdot)$ is Weibull with $\mu = .3$ and $c = .6$, we find that the MSE is minimized near $\tau = 13$ weeks, with relatively little change in the MSE over the range $\tau \in (10, 15)$ weeks.

1.2.3 Relationship Among Estimators

All of the estimators discussed in Sections 2.1 and 2.2 can be viewed as special cases of the most general maximum likelihood estimator of incidence given τ units of follow-up time,

$$\hat{\lambda} = \frac{n_{1\tau}}{N_1 \hat{p} \hat{\mu} \hat{H}(\tau)} \tag{1.21}$$

In general, it is not possible to estimate all of these quantities $\{p, \mu, H(\tau)\}$ well without external information or assumptions.

If we are in a setting where all apparent recent infections will eventually test positive on the less-sensitive test, and that the concept of "long-term non-progressor" does not apply, then $p = 1$, $n_1 = N_2$ and $\hat{H}(\tau) = n_{1\tau}/N_2$. Therefore,

$$\hat{\lambda} = \frac{n_{1\tau}}{N_1 \hat{p} \hat{\mu} \hat{H}(\tau)} = \frac{N_2}{N_1 \hat{\mu}}, \tag{1.22}$$

which represents Estimator A. Alternatively, we could assume that the value of p is known to be p_0, and not equal to 1. In this case we need to estimate $\{\mu, H(\tau)\}$, and results in Estimator C,

$$\hat{\lambda} = \frac{n_{1\tau}}{N_1 p_0 \hat{\mu} \hat{H}(\tau)}. \tag{1.23}$$

Rather than making assumptions about the value of p, one could instead make assumptions about $G(\cdot)$, and thus $h(\cdot)$, $H(\cdot)$, and μ. As discussed earlier in Wang and Lagakos (2010), if we have some knowledge of W^*, and choose $\tau \geq W^*$, then $\hat{H}(\tau) = 1$ and $n_1 = n_{1\tau}$, and

$$\hat{\lambda} = \frac{n_{1\tau}}{N_1 \hat{p} \hat{\mu} \hat{H}(\tau)} = \frac{n_1}{N_1 \hat{p} \hat{\mu}}, \tag{1.24}$$

which represents Estimator B. If we do not have any knowledge of W^*, or we are unable to follow subjects for $\tau > W^*$ for logistical reasons, the AF estimator may offer an attractive alternative if (μ, c) falls within the acceptable region. The AF estimator works well when

$G(\tau) \approx 0$. This results in $\hat{H}(\tau) = \tau/\hat{\mu}$ for small τ, and hence

$$\hat{\lambda} = \frac{n_{1\tau}}{N_1 \hat{p} \hat{\mu} \hat{H}(\tau)} = \frac{n_{1\tau}}{N_1 \hat{p} \tau}. \tag{1.25}$$

1.3 Simulation Studies

We have discussed 4 different estimators for estimating HIV incidence in an augmented cross-sectional study with follow-up time τ. We performed an extensive simulation study to compare their performance under a variety of situations. We assumed $(N, \lambda, \phi) = (6000, .02, .85)$ throughout. We let $p = 0.95$ or 1. We considered Weibull and Lognormal for the underlying distribution $G(\cdot)$, with mean $\mu = .3, .4$, or $.5$, and the coefficient of variation $c = .4$, or $.6$. The length of follow-up τ was taken to be 12, 24, or 52 weeks respectively. The frequency of follow-up visits is assumed to be biweekly, except for the AF estimator, which only requires a single follow-up visit at 12 weeks. In the case of Estimators A and C, we recorded right-censored observations using $b = 999$ weeks as the upper bound for the transition time of any patient whose transition is not observed during follow-up. For Estimator C, which calls for a pre-specified value for p, p_0 was taken to be .91, .95, or .99. For Estimators A, B, and C, we assumed a Weibull distribution in our maximization procedure throughout the paper because we found that using a Weibull distribution yields more robust results than using a Lognormal distribution, as in Wang and Lagakos (2010). Results corresponding to the usage of a Lognormal distribution in the maximization procedure are included in the Web Tables 2 and 3. We also considered confidence intervals based on profile likelihood methods. The width and coverage associated with profile likelihood confidence intervals are comparable to those obtained based on (1.10) (see Web Table 4), although we noticed that some computational difficulties arose when attempting to profile out the shape parameter of the Weibull distribution, forcing us to impose a lower bound on this value.

When $p = 1$, although the actual coverage of the confidence intervals based on Estimator A is close to the nominal level, we noticed that it tends to overestimate the true incidence rate

Table 1.1: *Simulation results for Estimators A and B, with N=6000, ϕ = .85 and λ=.02, and assuming Weibull $G(\cdot)$. $E(\hat{\lambda})$ and $sd(\hat{\lambda})$ denote average and standard deviation of estimates from 1000 simulated studies. $E(\hat{s})$ denotes average of likelihood-based estimates of the standard deviation from the 1000 experiments. Coverage denotes the proportion of experiments in which the true λ is contained in the nominal 95% confidence interval, and Width refers to the median width of the nominal 95% CI. Values in boldface represent scenarios where conditions for this estimator are met.*

Estimator A			p=1					p=0.95		
Weibull(μ=0.5,c=0.4)	E($\hat{\lambda}$)	sd($\hat{\lambda}$)	E($\hat{s}$)	Cov.	Width	E($\hat{\lambda}$)	sd($\hat{\lambda}$)	E($\hat{s}$)	Cov.	Width
τ=12 w, H(τ)=0.45	**0.026**	**0.0137**	**0.0137**	**0.981**	**0.028**	0.022	0.0078	0.0110	0.977	0.034
τ=24 w, H(τ)=0.80	**0.022**	**0.0060**	**0.0059**	**0.962**	**0.021**	0.023	0.0082	0.0099	0.959	0.035
τ=52 w, H(τ)=0.999	**0.020**	**0.0047**	**0.0044**	**0.934**	**0.017**	0.034	0.0128	0.0232	0.996	0.108
Weibull(μ=0.3,c=0.6)										
τ=12 w, H(τ)=0.64	**0.026**	**0.0191**	**0.0175**	**0.971**	**0.038**	0.023	0.0108	0.0169	0.962	0.074
τ=24 w, H(τ)=0.92	**0.022**	**0.0101**	**0.0090**	**0.946**	**0.030**	0.026	0.0127	0.0208	0.991	0.097
τ=52 w, H(τ) > 0.999	**0.021**	**0.0077**	**0.0073**	**0.931**	**0.026**	0.029	0.0190	0.0295	0.991	0.150
Lognorm(μ=0.5,c=0.4)										
τ=12 w, H(τ)=0.46	0.026	0.0140	0.0126	0.977	0.026	0.022	0.0075	0.0107	0.978	0.031
τ=24 w, H(τ)=0.81	0.022	0.0058	0.0057	0.949	0.021	0.023	0.0082	0.0099	0.968	0.034
τ=52 w, H(τ)=0.992	0.022	0.0052	0.0053	0.940	0.020	0.035	0.0133	0.0247	0.997	0.117
Lognorm(μ=0.3,c=0.6)										
τ=12 w, H(τ)=0.67	0.026	0.0199	0.0158	0.982	0.037	0.023	0.0105	0.0165	0.968	0.072
τ=24 w, H(τ)=0.92	0.025	0.0108	0.0107	0.971	0.036	0.027	0.0129	0.0229	0.996	0.115
τ=52 w, H(τ)=0.995	0.024	0.0097	0.0099	0.957	0.036	0.030	0.0187	0.0308	0.992	0.173
Estimator B										
Weibull(μ=0.5,c=0.4)	E($\hat{\lambda}$)	sd($\hat{\lambda}$)	E($\hat{s}$)	Cov.	Width	E($\hat{\lambda}$)	sd($\hat{\lambda}$)	E($\hat{s}$)	Cov.	Width
τ=12 w, H(τ)=0.45	0.022	0.0051	0.0059	0.960	0.023	0.022	0.0053	0.0061	0.963	0.024
τ=24 w, H(τ)=0.80	0.020	0.0037	0.0036	0.963	0.014	0.020	0.0038	0.0037	0.965	0.014
τ=52 w, H(τ)=0.999	**0.020**	**0.0045**	**0.0043**	**0.933**	**0.016**	**0.020**	**0.0045**	**0.0044**	**0.933**	**0.017**
Weibull(μ=0.3,c=0.6)										
τ=12 w, H(τ)=0.64	0.020	0.0060	0.0060	0.980	0.022	0.020	0.0064	0.0065	0.980	0.022
τ=24 w, H(τ)=0.92	0.019	0.0065	0.0054	0.861	0.018	0.019	0.0066	0.0057	0.842	0.018
τ=52 w, H(τ) > 0.999	**0.021**	**0.0076**	**0.0072**	**0.932**	**0.026**	**0.021**	**0.0079**	**0.0075**	**0.910**	**0.026**
Lognorm(μ=0.5,c=0.4)										
τ=12 w, H(τ)=0.46	0.022	0.0051	0.0059	0.941	0.023	0.022	0.0054	0.0060	0.954	0.024
τ=24 w, H(τ)=0.81	0.020	0.0037	0.0037	0.952	0.014	0.020	0.0039	0.0038	0.962	0.014
τ=52 w, H(τ)=0.992	0.021	0.0048	0.0048	0.943	0.019	0.021	0.0048	0.0049	0.951	0.019
Lognorm(μ=0.3,c=0.6)										
τ=12 w, H(τ)=0.67	0.021	0.0060	0.0061	0.975	0.022	0.021	0.0061	0.0064	0.985	0.023
τ=24 w, H(τ)=0.92	0.020	0.0068	0.0061	0.897	0.021	0.020	0.0064	0.0060	0.911	0.021
τ=52 w, H(τ)=0.995	0.023	0.0085	0.0089	0.958	0.033	0.023	0.0081	0.0088	0.968	0.032

in the settings we examined. For smaller value of τ, the variance associated with Estimator A is quite large in relative to the true incidence rate. Because λ is bounded below at 0, large variation in $\hat{\lambda}$ makes it more likely for us to obtain larger values of $\hat{\lambda}$ than smaller values of $\hat{\lambda}$, resulting in the average of $\hat{\lambda}$ to be biased high. We believe that this is a finite sample problem. We also performed additional simulations with larger sample size, and noticed that this overestimating phenomenon goes away. The precision of Estimator A improves for larger values of τ. When p=.95, the bias associated with Estimator A increases with larger τ and becomes substantial when $\tau = 52$ weeks. Moreover, its precision also worsens substantially as τ gets larger. These results suggest that Estimator A is very sensitive to assumptions about p. Unless we are certain about $p = 1$, the use of Estimator A would best be avoided.

When $H(\tau) \approx 1$, Estimator B performs well as expected, for Weibull or Lognormal $G(\cdot)$, as in Wang and Lagakos (2010). Its actual coverage level is generally near the nominal level. When $H(\tau) < 1$, we observe modest bias for Estimator B and the actual coverage level can be substantially lower than the nominal level when $\tau = 24$ weeks. Interestingly, we find that Estimator B provides adequate coverage when $\tau = 12$ weeks, and showed greater precision than Estimator A in such scenarios. This phenomenon is closely related to the good performance of the AF estimator and was also part of our motivation for considering the AF estimator.

When p_0 is correctly specified, Estimator C performs well in general, showing coverage near the nominal level and precision improving with increasing values of τ. When the hypothesized p_0 is smaller than the true p, Estimator C behaves like Estimator B, maintaining actual coverage greater than 92% for $\tau = 12$ or $\tau = 52$ weeks and can have the under-coverage problem for $\tau = 24$ weeks in some settings. When the hypothesized p_0 is larger than the true p, Estimator C behaves like Estimator A, and can lead to substantial bias and greater variances.

As the follow-up time increases, the AF estimator becomes less variable yet more biased. The 12-week AF Estimator performs well in all scenarios (see Table , with further simulation

Table 1.2: *Simulation results for Estimator C, with N=6,000, ϕ = .85 and λ=0.02, and assuming Weibull $G(\cdot)$. $E(\hat{\lambda})$ and $sd(\hat{\lambda})$ denote average and standard deviation of estimates from 1000 simulated studies. $E(\hat{s})$ denotes average of likelihood-based estimates of the standard deviation from the 1000 simulations. Coverage denotes the proportion of simulations in which the true λ is contained in the nominal 95% confidence interval, and Width refers to the median width of the nominal 95% CI. Values in boldface represent scenarios where conditions for this estimator are met.*

Estimator C			p=1					p=0.95		
Weibull(μ=0.5,c=0.4)	E($\hat{\lambda}$)	sd($\hat{\lambda}$)	E($\hat{s}$)	Cov.	Width	E($\hat{\lambda}$)	sd($\hat{\lambda}$)	E($\hat{s}$)	Cov.	Width
p_o=0.95										
τ=12 w, H(τ)=0.45	0.022	0.0056	0.0060	0.971	0.022	**0.025**	**0.0157**	**0.0138**	**0.992**	**0.028**
τ=24 w, H(τ)=0.80	0.019	0.0037	0.0036	0.960	0.013	**0.022**	**0.0075**	**0.0062**	**0.963**	**0.020**
τ=52 w, H(τ)=0.999	0.020	0.0043	0.0042	0.936	0.016	**0.021**	**0.0060**	**0.0051**	**0.927**	**0.018**
p_o=0.99										
τ=12 w, H(τ)=0.45	0.025	0.0149	0.0124	0.994	0.027	0.023	0.0090	0.0121	0.998	0.031
τ=24 w, H(τ)=0.80	0.020	0.0043	0.0042	0.964	0.014	0.023	0.0085	0.0100	0.975	0.034
τ=52 w, H(τ)=0.999	0.020	0.0043	0.0042	0.936	0.016	0.038	0.0152	0.0263	0.989	0.121
Weibull(μ=0.3,c=0.6)										
p_o=0.95										
τ=12 w, H(τ)=0.64	0.020	0.0071	0.0064	0.984	0.021	**0.027**	**0.0228**	**0.0210**	**0.989**	**0.036**
τ=24 w, H(τ)=0.92	0.019	0.0066	0.0054	0.843	0.018	**0.024**	**0.0161**	**0.0132**	**0.932**	**0.030**
τ=52 w, H(τ) > 0.999	0.021	0.0077	0.0073	0.924	0.027	**0.022**	**0.0150**	**0.0099**	**0.926**	**0.027**
p_o=0.99										
τ=12 w, H(τ)=0.64	0.023	0.0165	0.0107	0.981	0.023	0.025	0.0131	0.0198	0.999	0.078
τ=24 w, H(τ)=0.92	0.019	0.0067	0.0056	0.850	0.019	0.029	0.0166	0.0251	0.990	0.116
τ=52 w, H(τ) > 0.999	0.021	0.0077	0.0073	0.924	0.027	0.042	0.0302	0.0505	0.999	0.296
Lognorm(μ=0.5,c=0.4)										
p_o=0.95										
τ=12 w, H(τ)=0.46	0.023	0.0053	0.0059	0.950	0.022	0.026	0.0133	0.0136	0.985	0.027
τ=24 w, H(τ)=0.81	0.020	0.0038	0.0037	0.955	0.014	0.022	0.0059	0.0058	0.969	0.019
τ=52 w, H(τ)=0.992	0.021	0.0046	0.0048	0.954	0.019	0.022	0.0055	0.0055	0.948	0.020
p_o=0.99										
τ=12 w, H(τ)=0.46	0.025	0.0120	0.0110	0.978	0.025	0.023	0.0088	0.0120	0.996	0.031
τ=24 w, H(τ)=0.81	0.021	0.0044	0.0042	0.950	0.015	0.024	0.0087	0.0099	0.994	0.033
τ=52 w, H(τ)=0.992	0.021	0.0046	0.0048	0.954	0.019	0.038	0.0157	0.0279	0.988	0.132
Lognorm(μ=0.3,c=0.6)										
p_o=0.95										
τ=12 w, H(τ)=0.67	0.021	0.0059	0.0059	0.972	0.022	0.026	0.0192	0.0170	0.992	0.034
τ=24 w, H(τ)=0.92	0.020	0.0064	0.0061	0.901	0.022	0.025	0.0139	0.0123	0.962	0.033
τ=52 w, H(τ)=0.995	0.023	0.0080	0.0088	0.965	0.033	0.025	0.0143	0.0117	0.960	0.035
p_o=0.99										
τ=12 w, H(τ)=0.67	0.022	0.0094	0.0081	0.971	0.023	0.025	0.0137	0.0194	1.000	0.071
τ=24 w, H(τ)=0.92	0.020	0.0066	0.0063	0.910	0.023	0.031	0.0167	0.0284	1.000	0.144
τ=52 w, H(τ)=0.995	0.023	0.0080	0.0088	0.965	0.033	0.046	0.0357	0.0573	0.999	0.351

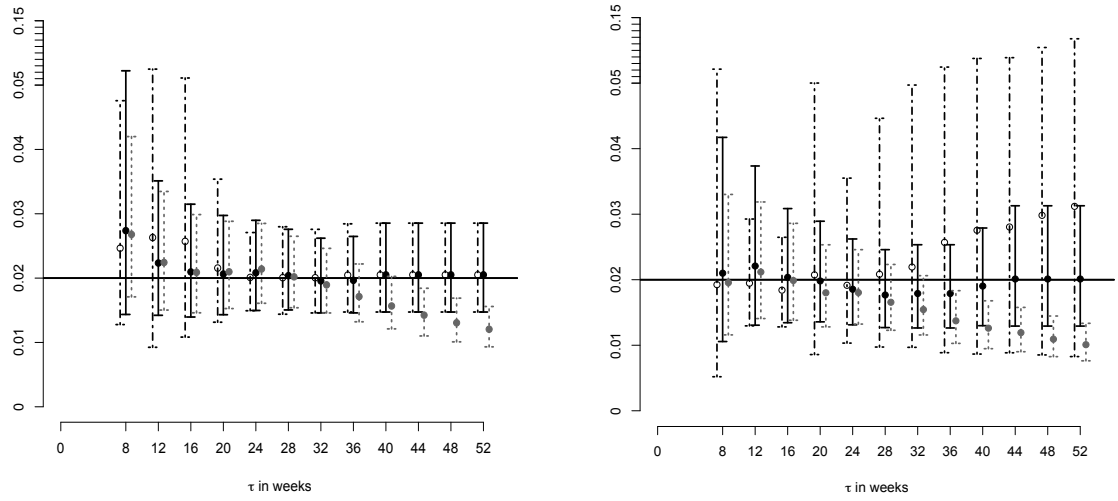
results in Web Table 5) within the acceptable region outlined in 1.2(a). For these situations, it shows greater precision than all other estimators with $\tau = 12$ weeks. For settings in the non-acceptable region, the use of the AF estimator should be avoided. For example, when $(\mu, c) = (.3, .8)$, the actual coverage level of the AF estimator drops to 85%.

Table 1.3: *Simulation results for the AF Estimator, with N=6000, ϕ = .85 and λ=0.02. E($\hat{\lambda}$) and sd($\hat{\lambda}$) denote average and standard deviation of estimates from 1,000 simulated studies. E($\hat{s}$) denotes average of likelihood-based estimates of the standard deviation from the 1,000 simulations. Coverage denotes the proportion of simulations in which the true λ is contained in the nominal 95% confidence interval, and Width refers to the median width of the nominal 95% CI.*

AF Estimator			p=1					p=0.95		
Weibull(μ=0.5,c=0.4)	E($\hat{\lambda}$)	sd($\hat{\lambda}$)	E($\hat{s}$)	Cov.	Width	E($\hat{\lambda}$)	sd($\hat{\lambda}$)	E($\hat{s}$)	Cov.	Width
τ=12 w, H(τ)=0.45	0.020	0.0043	0.0042	0.945	0.017	0.021	0.0042	0.0043	0.951	0.018
τ=24 w, H(τ)=0.80	0.018	0.0028	0.0027	0.886	0.011	0.018	0.0028	0.0028	0.887	0.011
τ=52 w, H(τ)=0.999	0.010	0.0015	0.0014	0.000	0.006	0.010	0.0015	0.0014	0.000	0.006
Weibull(μ=0.3,c=0.6)										
τ=12 w, H(τ)=0.64	0.017	0.0039	0.0038	0.915	0.015	0.017	0.0039	0.0039	0.923	0.016
τ=24 w, H(τ)=0.92	0.012	0.0022	0.0023	0.158	0.009	0.012	0.0024	0.0023	0.161	0.009
τ=52 w, H(τ) > 0.999	0.006	0.0011	0.0011	0.000	0.004	0.006	0.0012	0.0011	0.000	0.004
Lognorm(μ=0.5,c=0.4)										
τ=12 w, H(τ)=0.46	0.021	0.0041	0.0042	0.952	0.017	0.021	0.0045	0.0043	0.935	0.017
τ=24 w, H(τ)=0.81	0.018	0.0028	0.0027	0.887	0.011	0.018	0.0028	0.0028	0.899	0.011
τ=52 w, H(τ)=0.992	0.010	0.0014	0.0014	0.000	0.006	0.010	0.0014	0.0014	0.000	0.006
Lognorm(μ=0.3,c=0.6)										
τ=12 w, H(τ)=0.67	0.018	0.0039	0.0039	0.930	0.016	0.018	0.0040	0.0040	0.944	0.016
τ=24 w, H(τ)=0.92	0.012	0.0023	0.0022	0.170	0.009	0.012	0.0023	0.0023	0.180	0.009
τ=52 w, H(τ)=0.995	0.006	0.0011	0.0011	0.000	0.004	0.006	0.0011	0.0011	0.000	0.004

Illustrative Examples Because data following such an augmented cross-sectional design do not yet exist, we chose 2 of our simulated data sets from the scenario with Weibull $G(\cdot)$ and $(\mu, c) = (.5, .4)$ for illustration. The first data set was generated with p=1, indicating that the conditions for Estimator A were met for any choice of τ, while the assumptions underlying Estimator B were approximately satisfied ($H(\tau) > .99$) for $\tau \geq$ 44 weeks. In Figure 1.3(a), we see that both incidence estimators have point estimates near the true incidence value and confidence intervals containing the true value, for any choice of τ. Furthermore, we see that the width of the confidence interval of Estimator B is comparable to that of Estimator A for most values of τ. The AF Estimator is shown in gray and yields results similar to those of the other estimators for smaller values of τ. The second data set was generated with p=.95, indicating that the conditions for Estimator

A were never met in this setting. In this example (Fig. 1.3(b)), we see that Estimator A suffers greatly in its precision as τ increases. The 12-week AF Estimator yields results similar to those of Estimator B for $\tau = 12$ weeks, though with a somewhat smaller interval width.



(a) p=1 (b) p=0.95

Figure 1.3: *Point estimates and 95% confidence intervals for the incidence rate at different stopping times τ. The data sets were chosen from the simulation study described in Section 3.*
Black Dashed: Estimator A, Black Solid: Estimator B, Gray Dotted: AF Estimator
$G(\cdot) \sim$ Weibull, (μ,c) = (.5,.4)

1.4 Discussion

In this paper, we assess how the performance of various estimators are affected by length of follow-up τ and/or assumptions about the size of the "long-term non-progressor" population. We found that Estimator A, which assumes $p = 1$, can result in substantial bias and great loss of precision if the true p is less than 1. This pattern also holds for Estimator C when we overestimate the true p. When p is correctly specified, both Estimator A and Estimator C work well regardless of length of follow-up. However, p is often not known in practice and

is expected to vary according to the type of less-sensitive test and population of interest. Estimator B does not require prior knowledge of p, but instead requires length of follow-up to be long enough so that subjects who have not become reactive to the less-sensitive test at the end of the follow-up will remain non-reactive permanently. We found that Estimator B is robust to modest violation of this assumption. Interestingly, when we apply Estimator B with very short follow-up, say, 12 weeks, it works reasonably well and has led us to consider a new HIV incidence estimator based on an abbreviated follow-up time, the AF estimator.

The AF estimator has the benefit of not requiring any specific external estimate of p or μ, as well as substantially shortening the amount of follow-up compared to earlier augmented cross-sectional designs without substantial loss of precision. The AF estimator requires rather general assumptions about the distribution of the time spent in the "recent infection" state, and its use will become more feasible as more information becomes available characterizing these distributions associated with particular less-sensitive tests in different populations. We note that the length of follow-up τ may be chosen to be somewhat smaller than 12 weeks, which will further relax the distributional assumptions necessary for the AF estimator. However, the sample size will need to be increased accordingly, as shown in (1.20), to maintain desired precision. Since the precision of the AF estimator is dependent solely on the observed number of transitions into State 3, it is recommended that all N_2 patients be retested at the 12-week follow-up visit. However, if it is only feasible to follow a random subsample of size $n_2 < N_2$ of these patients, the AF estimator can be easily adapted by replacing the value $n_{1\tau}$ with $n_{1\tau} \cdot N_2/n_2$ in (1.12) and (1.13). The procedure for calculating the confidence intervals remains unchanged.

The AF estimator, despite its attractive properties, is biased and therefore currently can only serve as a crude estimator when resources are limited. When possible, we recommend the use of Estimator B proposed in Wang and Lagakos (2010) and to follow subjects for a reasonably long period of time. To facilitate the implementation of the augmented cross-sectional design and the use of Estimator B, it would be helpful to provide guidance on several design issues such as sample size of the cross-sectional sample, number of subjects

needed to be followed and visit frequency. Because an explicit variance formula for $\hat{\mu}$ is not readily attainable, we do not have a formula for sample size yet. Using simulations and a grid search, we can calculate sample sizes needed to estimate incidence in a population of interest. For example, for one of the settings we examined in 2.2.3, when $(\phi, \lambda, p) = (.75, .04, .94)$, if we assume the underlying distribution for the window period is Weibull with $\mu = .5$ and $c = .3$, we would need about 2200 subjects in the cross-sectional sample and to follow all subjects found to be in 'recent state' for about 1 year, in order for $\delta = 2$ with probability 80%. Interestingly, we found that the changes in the required sample size were negligible when we relaxed a 2-week visit schedule to a 6-week visit schedule. Whether this holds true in general warrants further research.

We emphasize the importance of quantifying p. With correct knowledge of p, Estimator A or Estimator C would work well with shorter follow-up than that required by Estimator B. Furthermore, if p is known, and it is possible to devise a less-sensitive test which has the property that subjects will not become reactive for a minimum period of time τ^*, then the AF estimator would be unbiased with $\tau < \tau^*$.

We investigated the implementation of a fully Bayesian model to address the problem at hand with what we consider to be fairly vague priors (Web Appendix B). The results are included in Web Table 5. When the true parameters lie well within the interior of the prior distributions, the Bayesian method produces credible intervals with width narrower than those obtained from Estimators A, B, and C, and comparable to those from the AF estimator. However, the Bayesian method is seen to have under-coverage problems when the true parameters are close to the boundary of their prior distributions. For example, when true μ=.3 years and the prior distribution was chosen as Uniform on the interval from .25 years to 1.5 years, even though the prior distribution has non-negligible mass near the true value, the posterior distribution is not symmetric around the true value, due to the truncation at .25 years. The strong negative correlation between $\hat{\lambda}$ and $\hat{\mu}$ can result in $\hat{\lambda}$ being biased low. Thus the corresponding 95% credible interval may not have the nominal coverage. It would also be useful to investigate whether the EM algorithm can be used effectively in the

calculation of Estimator C, by iteratively assigning the $n_{0\tau}$ patients to either State 2 or State 4, based on the estimated parameter values at each step, then re-estimating the parameter values by maximizing the updated likelihood function, based on the state assignments.

Recent evidence suggests that use of antiretroviral treatments (ART) can affect the results of less-sensitive diagnostic tests (Marinda et al., 2010). For this reason, it is critical that patients' ART use be documented. One suggestion is to exclude patients with known ART use from the initial cross-sectional sample (Mcdougal et al., 2006), though this could potentially impact the validity of the resulting prevalence estimate if ART use is common in the study region. Suppose that some number of the N_2 infected patients in the cross-sectional survey who test negative on the less-sensitive test are known to be receiving ART. Assuming that these patients' test results do not change during the follow-up period, all n^* patients could be reassigned to State 3 or "non-recent infection". An alternative would be to reassign $\hat{p}n^*$ of these patients to State 3, while keeping the remaining $(1-\hat{p})n^*$ in State 4. Both methods could be employed and the resulting estimates compared as a sort of sensitivity analysis. The assumption being made here is that the population of patients currently receiving ART at a given time is likely to be comprised of a disproportionately smaller percentage of "long-term non-progressors" than the overall infected population.

Supplementary Materials

Web Appendices and Tables referenced in Sections 1.2.1, 1.2.2, 1.3, and 1.4 are available under the Paper Information link at the Biometrics website `http://www.biometrics.tibs.org`.

Acknowledgements

We thank the editor, an associate editor and two referees for their comments, which improved the paper. This research was supported by grants R37 AI24643 and T32 AI007358 from the National Institutes of Health.

1.5 Appendix

1.5.1 Appendix 1: Likelihood Function for Model in Section 1.2.1

The state prevalence probabilities at time t are given by (Balasubramanian and Lagakos, 2010; Wang and Lagakos, 2009)

$$\pi_1(t) \stackrel{def}{=} P(\text{in State 1 at time } t) = 1 - F(t) = \phi, \tag{1.26}$$

$$\pi_2(t) \stackrel{def}{=} P(\text{in State 2 at time } t) = p\phi\lambda\mu, \tag{1.27}$$

$$\pi_3(t) \stackrel{def}{=} P(\text{in State 3 at time } t) = p(1 - \phi - \phi\lambda\mu), \tag{1.28}$$

$$\pi_4(t) \stackrel{def}{=} P(\text{in State 4 at time } t) = (1 - p)(1 - \phi). \tag{1.29}$$

The cumulative distribution function of X, for $x > 0$, conditional on t, is (Wang and Lagakos, 2010):

$$H(x \mid t) = \int_0^x \frac{1 - G(v)}{\mu} dv. \tag{1.30}$$

In our setting, only those who are observed to test positive on the less-sensitive test within the τ units of follow-up (i.e. those with forward recurrence time $X < \tau$), will be determined to definitely be in State 2 at time t. Thus the probability of actually being in State 2 and being observed to transition to State 3 during the follow-up time is $\pi_2 H(\tau \mid t) = \phi\lambda p\mu H(\tau \mid t)$. The probability of actually being in State 2 at the time of the cross-sectional

sample, but not being observed to transition into State 3 during the follow-up period is then

$$\pi_2\{1 - H(\tau \mid t)\} = \phi\lambda p\mu\{1 - H(\tau \mid t)\}. \tag{1.31}$$

The additional information from the interval-censored forward recurrence times of those patients determined to be in State 2 is

$$P(a < x < b | x < \tau) = \frac{H(b \mid t) - H(a \mid t)}{H(\tau \mid t)}. \tag{1.32}$$

1.5.2 Appendix 2: Probability Limit of $\hat{\lambda}_B$

$\hat{\lambda}_B = \frac{n_{1\tau}}{N_1 \hat{p}_\tau \hat{\mu}_\tau}$, where $\hat{p}_\tau = \frac{N_3 + n_{1\tau}}{N_3 + n_{1\tau} + n_{0\tau}} = \frac{N_3 + n_{1\tau}}{N_3 + N_2}$. Let μ_τ be the probability limit of $\hat{\mu}_\tau$. From

$$\frac{n_{1\tau}}{N} \xrightarrow{p} p\phi\lambda\mu H(\tau), \frac{N_1}{N} \xrightarrow{p} \phi, \frac{N_2 + N_3}{N} \xrightarrow{p} (1 - \phi), \text{ and } \frac{N_3}{N} \xrightarrow{p} p(1 - \phi - \phi\lambda\mu), \tag{1.33}$$

we have

$$\hat{p}_\tau = \frac{(N_3 + n_{1\tau})/N}{(N_3 + N_2)/N} \xrightarrow{p} \frac{p(1 - \phi - \phi\lambda\mu) + p\phi\lambda\mu H(\tau)}{1 - \phi} = \frac{p[1 - \phi - \phi\lambda\mu\{1 - H(\tau)\}]}{1 - \phi}, \tag{1.34}$$

and

$$\hat{\lambda}_B = \frac{n_{1\tau}/N}{N_1/N\hat{p}_\tau\hat{\mu}_\tau} \xrightarrow{p} \frac{p\phi\lambda\mu H(\tau)}{\phi\frac{p[1-\phi-\phi\lambda\mu\{1-H(\tau)\}]}{1-\phi}\mu_\tau} = \lambda\frac{\mu}{\mu_\tau}\frac{(1-\phi)H(\tau)}{[1 - \phi - \phi\lambda\mu\{1 - H(\tau)\}]}. \tag{1.35}$$

1.5.3 Appendix 3: Relationship of $f(t)$ and $f_2(t)$

$$\begin{aligned} f_2(t) &= P(\text{transition from State 2 to State 3 at time } t) \\ &= \int P(\text{infected at time } t-u) \times P(\text{in subgroup that will transition to State 3}) \\ &\quad \times P(\text{sojourn time in State 2} = u | \text{in subgroup that will transition to State 3})\ du. \end{aligned} \tag{1.36}$$

Under the assumption that $f(u)$ is constant over $u \in (t - W^*, t)$,

$$f_2(t) = \int_{u=0}^{W^*} f(t-u) \cdot p \cdot g(u) du = f \cdot p \int_{u=0}^{W^*} g(u) du = f \cdot p \cdot G(W^*) = pf \tag{1.37}$$

1.5.4 Appendix 4: Expectation of AF Estimator

Recall $\hat{\lambda}_{AF} = \frac{n_{1\tau}}{N_1 \hat{p}_\tau \tau} = \frac{n_{1\tau}(N_3+N_2)}{N_1 \tau (N_3+n_{1\tau})} = \frac{N_3+N_2}{N_1} \frac{n_{1\tau}}{N_3+n_{1\tau}} \frac{1}{\tau}$, therefore,

$$E(\hat{\lambda}_{AF}) = E(\frac{N_3+N_2}{N_1} \frac{n_{1\tau}}{N_3+n_{1\tau}} \frac{1}{\tau}) = \frac{1}{\tau} E(\frac{N_3+N_2}{N_1} \frac{n_{1\tau}}{N_3+n_{1\tau}}). \tag{1.38}$$

Because $(N_1, n_{1\tau}, n_{0\tau}, N_3) \sim \text{Multinomial}(N; p_1, p_2, p_3, p_4)$, we have

$$n_{1\tau}|(N_1, n_{0\tau}) \sim \text{Binomial}(N - N_1 - n_{0\tau}, \frac{p_2}{1-p_1-p_3}). \tag{1.39}$$

Now, we want to show E(XY) = E(X) E(Y), where $X = \frac{N_3+N_2}{N_1}$, $Y = \frac{n_{1\tau}}{N_3+n_{1\tau}}$.

$$\begin{aligned} E(Y) &= E\{E(Y|N_1, n_{0\tau}))\} = E\{E(\frac{n_{1\tau}}{n_{1\tau}+N_3}|N_1, n_{0\tau})\} = E\{\frac{1}{n_{1\tau}+N_3} E(n_{1\tau}|N_1, n_{0\tau})\} \\ &= E(\frac{1}{n_{1\tau}+N_3} \frac{(n_{1\tau}+N_3)p_2}{p_2+p_4}) = E(\frac{p_2}{p_2+p_4}) = \frac{p_2}{p_2+p_4} \end{aligned} \tag{1.40}$$

and

$$\text{E}(Y|X) = E\{E(Y|N_1, n_{0\tau})|X\} = E(\frac{p_2}{p_2+p_4}|X) = \frac{p_2}{p_2+p_4} \tag{1.41}$$

Because $E(Y|X) = E(Y) \Rightarrow E(XY) = E(X)E(Y)$,

$$\begin{aligned} \text{so} \qquad E(\hat{\lambda}_{AF}) &= \frac{1}{\tau} E(\frac{N_3+N_2}{N_1}) E(\frac{n_{1\tau}}{N_3+n_{1\tau}}) \\ &\approx \frac{1}{\tau} \frac{1-\phi}{\phi} \frac{p\lambda\phi\mu H(\tau)}{1-\phi-p\phi\lambda\mu\{1-H(\tau)\}} \\ &= \frac{\lambda\mu H(\tau)}{\tau} \frac{1}{1-p_r\{1-H(\tau)\}} \end{aligned} \tag{1.42}$$

$$\text{where} \qquad p_r = \frac{\lambda\phi\mu}{1-\phi} = P(\text{recent infection}|\text{currently infected}). \tag{1.43}$$

1.5.5 Appendix 5: Approximate Variance of AF Estimator

Using (1.12) and (1.13), we have

$$log(\hat{\lambda}_{AF}) = log(\frac{N_2 + N_3}{N_1}) + log(\frac{n_{1\tau}}{N_3 + n_{1\tau}}) - log(\tau) \tag{1.44}$$

Because the two terms are shown to be independent in Appendix 4, this leads to

$$Var\{log(\hat{\lambda}_{AF})\} = Var\{log(\frac{N_2 + N_3}{N_1})\} + Var\{log(\frac{n_{1\tau}}{N_3 + n_{1\tau}})\} \tag{1.45}$$

Note that $log\{(N_2 + N_3)/N_1\}$ is the log odds of HIV prevalence, its variance can be consistently estimated by $1/N_1 + 1/(N_2 + N_3)$ (Woolf et al., 1955). Similarly, by application of the δ-method and common variance formula for sample proportion, the variance of $log\{n_{1\tau}/(N_3 + n_{1\tau})\}$ can be approximated by $N_3/\{n_{1\tau}(N_3 + n_{1\tau})\}$. Therefore,

$$\widehat{Var}\{log(\hat{\lambda}_{AF})\} \approx \frac{1}{N_1} + \frac{1}{N_2 + N_3} + \frac{N_3}{n_{1\tau}(N_3 + n_{1\tau})} \approx \frac{1}{n_{1\tau}}, \tag{1.46}$$

where the last approximation follows from that N_1 and N_3 are usually much larger than $n_{1\tau}$, and therefore $1/N_1$ and $1/(N_2+N_3)$ are negligible compared to $1/n_{1\tau}$, and $N_3/(N_3+n_{1\tau}) \approx 1$.

1.5.6 Appendix 6: Bounding the Bias of AF Estimator

Suppose we wish to bound the bias of the AF Estimator, such that the expected value of the AF Estimator is close to the true incidence rate, to within a factor of B. That is, $B\lambda \leq E(\hat{\lambda}_{AF}) \leq \lambda/B$. Using (1.42), after some simplification we have

$$B\tau[1 - p_r\{1 - H(\tau)\}] \leq \int_0^\tau 1 - G(x)\ dx \leq \frac{\tau}{B}[1 - p_r\{1 - H(\tau)\}] \tag{1.47}$$

where p_r is defined in (1.43).

If the desired bound B is chosen such that $B \leq (1 - p_r)$, then the second inequality is always satisfied, because $\tau \leq \tau[1 - p_r\{1 - H(\tau)\}]/B$ and $\int_0^\tau 1 - G(x)\ dx \leq \tau$ since $G(\cdot) \geq 0$

always. In this case, we only need to consider the first inequality, which can be expressed as

$$\tau \leq \frac{\int_0^\tau 1 - G(x)\ dx}{B[1 - p_r\{1 - H(\tau)\}]}. \tag{1.48}$$

If a tighter bound B is desired, such that $B > (1 - p_r)$, we must also satisfy the second inequality, which can be expressed as

$$\tau \geq B\frac{\int_0^\tau 1 - G(x)dx}{[1 - p_r\{1 - H(\tau)\}]}. \tag{1.49}$$

Estimating Subject-Specific Treatment Differences for Risk-Benefit Assessment with Applications to Beta-Blocker Effectiveness Trials

Brian Claggett[1], Lihui Zhao[2], Lu Tian[3], Davide Castagno[4], and Lee-Jen Wei[1]

[1]Department of Biostatistics

Harvard School of Public Health

[2]Department of Preventive Medicine

Northwestern University

[3]Department of Health Research and Policy

Stanford University School of Medicine

[4]Department of Internal Medicine

University of Turin

2.1 Introduction

Consider a randomized, comparative clinical trial in which a treatment is assessed against a control with respect to their risk-benefit profiles. Conventionally, a single treatment contrast is utilized to assess an overall treatment difference with respect to efficacy, in addition to a global measure of toxicity, over a rather heterogeneous population. Unfortunately, the resulting inference about these two measures are rather difficult to interpret in clinical practice. Making patient-specific decisions based on estimated population-averaged effects can lead to sub-optimal patient care (Kent and Hayward, 2007). A positive (negative) trial based on these two overall measures does not mean that every future patient should (should not) be treated by the new treatment. To bring the clinical trial results to the patient's bedside, we may utilize the patient's characteristics which relate to the response variable to perform so-called personalized or stratified medicine. Unfortunately, the typical ad hoc subgroup analysis of clinical studies is not credible (Wang et al., 2007). Moreover, such subgroup analysis is often conducted by investigating the effect of only a single predictor at a time and therefore may not be effective in identifying patients who would benefit from the new treatment.

In this paper, we present a systematic approach to nonparametrically estimate subject-specific treatment differences from a risk-benefit perspective. To address the issue of risk-benefit assessment, we propose categorizing each patient into one of several ordered categories at a given follow-up time, more thoroughly reflecting the patient's treatment experience. That is, for each study subject, the observations are times to events that could be used define the efficacy and toxicity of the treatment, which are then used to categorize a patient's overall clinical outcome. In the event of censoring of one or more of these event outcomes, we use inverse-probability weighting to obtain consistent estimates of the associated probabilities. In order to assess these treatment effects nonparametrically from a personalized medicine perspective, we use a training data set to first build a parametric univariate scoring system using baseline variables and then stratify subjects in the target data set accordingly. We then

estimate the treatment effects nonparametrically with respect to the risk-benefit categories.

When there is a single baseline covariate involved, Bonetti et al. (2000); Song and Pepe (2004) and Bonetti and Gelber (2004) have proposed novel statistical procedures for identifying a subgroup of patients who would benefit from the new treatment with respect to efficacy. A recent paper by Janes et al. (2011), which is based on previous work by Pepe (2004), Huang et al. (2007), and Pepe et al. (2008), provides practical guidelines for measuring the performance of individual markers for treatment selection. By incorporating more than one baseline covariate at a time, our approach is similar in spirit to Cai et al. (2011) and Li et al. (2011). However, they both used a single study to create a scoring system by fitting a prespecified model without model evaluation or variable selection. They then use the same data set to make inferences for either the treatment difference without considering multiple within-patient outcomes or for risk predictions for a single treatment group only.

2.2 Data

2.2.1 Target Data

Our data set of interest comes from a recent clinical trial, "Beta-Blocker Evaluation of Survival Trial" (BEST), which compared a beta-blocker to placebo in patients with heart failure (HF), with a primary endpoint of all-cause mortality. In this trial, other monitored patient outcomes included timing of hospitalizations, with cause of each hospitalization recorded as being due to the patient's chronic heart failure or for other reasons, and all deaths were adjudicated as being due to cardiovascular causes (CV death), or otherwise (non-CV death) (Beta-Blocker Evaluation of Survival Trial Investigators, 2001). Note that in the BEST trial, the non-fatal event times were not censored by other non-fatal event times. This trial enrolled 2708 patients and is of particular interest because of the observed marginally significant treatment effect (p=0.10), with an estimated hazard ratio of 0.90 (0.78, 1.02) for the

effect of treatment on mortality, the primary endpoint of the study. This result stands in contrast to other studies of beta-blockers in patients with heart failure which were conducted around the same time, each of which showed highly significant beneficial treatment effects with hazard ratios ≤ 0.67 and p-values ≤ 0.001 (Domanski et al., 2003).

2.2.2 Training Data

The usage of an independent training data set is crucial for multiple reasons. Primarily, the proposed training/validation approach allows one to avoid the nontrivial "self-serving" bias that can result from performing the model-building and variable selection process, creating the score, stratifying subjects, and estimating subject-specific treatment differences all within the same data set. Practically, the medical literature contains multiple examples of beta-blocker trials which have suggested differential treatment effects according to various baseline covariates, none of which were able to be confirmed in subsequent trials. Furthermore, it has been noted that the validity of subgroup findings is strengthened when able to be validated prospectively (Friedman et al., 2010). Finally, since formal evaluations of new drugs or devices usually require two well-conducted studies, it is often feasible that a trial sponsor has access to existing trial data that is directly relevant to the study of interest. Because we have access to only this particular trial, we instead mimic a hypothetical interim analysis which could have been conducted by analyzing only the first 900 patients ($\sim 33\%$ of total enrollment).

We used the outcome data (mortality and hospitalizations) from these early patients, along with 16 available patient-level covariates to derive the parametric scoring systems which would be used to stratify the remaining patients in the BEST study.

2.2.3 Patient Outcome Categories

In order to both develop the scoring systems in the training data set and evaluate the patient outcomes in the validation set, a classification system is needed describe each patient's status at a fixed followup time t_0. A simple and intuitive classification scheme, as described below, would involve five mutually exclusive and exhaustive categories, depending on which, if any, events have been experienced by a patient prior to t_0. This classification scheme is designed to account for events whose reduction represents an anticipated "benefit" of treatment (i.e. CV hospitalization, CV death), as well as events which may plausibly occur more frequently as a result of treatment, and are therefore considered potential treatment "risks" (i.e. non-CV hospitalization, non-CV death).

For patient i receiving treatment j, let $T_{(1)ij}$ denote a patient's time to first non-CV hospitalization, $T_{(2)ij}$ denote the time to the first CV-related hospitalization, and $T_{(3)ij}$ denote the time to death, where $i = 1, ..., n_j; j = 1, 2$. Furthermore, let $\delta_{ij} = 1$ if a patient's death is classified as being CV-related, and $\delta_{ij} = 0$ otherwise. In the following analysis, we will target the joint distribution of $\{(T_{(1)j}, T_{(2)j}, T_{(3)j}, \delta_j), j = 1, 2\}$ for patients with given parametric score, which is needed for risk-benefit analysis on a personalized level. This novel approach is different from the conventional risk-benefit analysis, whose main objective is to separately characterize distribution of $(T_{(3)j}, \delta_j)$, as well as the marginal distributions of $T_{(1)j}$ and $T_{(2)j}$ over the entire population. To this end, let us define a patient's classification at t_0 by $\epsilon(t_0)_{ij}$, where

- $\epsilon(t_0) = 0$ if $t_0 < (T_{(1)} \wedge T_{(2)} \wedge T_{(3)})$ (i.e., patient is "alive and healthy")
- $\epsilon(t_0) = 1$ if $T_{(1)} \leq t_0 < (T_{(2)} \wedge T_{(3)})$ (i.e., patient is "alive and hospitalized without worsening HF")
- $\epsilon(t_0) = 2$ if $T_{(2)} \leq t_0 < T_{(3)}$ (i.e., patient is "alive with worsening HF")
- $\epsilon(t_0) = 3$ if $T_{(3)} \leq t_0, \delta = 0$ (i.e., patient experienced "non-CV death")
- $\epsilon(t_0) = 4$ if $T_{(3)} \leq t_0, \delta = 1$ (i.e., patient experienced "CV death")

This classification is obviously defined by $T_{(1)ij}, T_{(2)ij}, T_{(3)ij}$, and δ_{ij} jointly and its conditional distributions for patients with given parametric score in the treatment as well as control groups may guide us to make optimal treatment recommendations for this group of patients. For clarity of notation, we will assume t_0 is fixed and will henceforth denote classification status using ϵ instead.

2.3 Building a Scoring System via the Training Data Set

To begin, we use the BEST training data set to build a scoring system using the patients' baseline characteristics with respect to the above ordinal outcome categories. In order to use nonparametric methods in the evaluation stage which estimate the expected treatment effects over a univariate scoring system, the goal of the first stage of our procedure is to reduce the multivariate covariate vector U to a single covariate $S = f(U)$ which will be effective in grouping patients who would experience similar treatment effects. Thus, even though we are interested in assessing the effect of treatment on more than one outcome, we must construct a single score so that each patient i is associated with a single covariate value S_i. Specifically, for this training set, each subject was assigned to a particular treatment j, where $j = 1, 2$. Let U_j be the vector of baseline covariates, including the treatment indicator τ_j with $\tau_j = 1$ in the treated group, and 0 otherwise. Let C_j be the censoring variable, which is assumed to be independent of U_j and all T_j. Furthermore, let $X_{(1)j} = \min(T_{(1)j}, C_j)$, $X_{(2)j} = \min(T_{(2)j}, C_j)$, and $X_{(3)j} = \min(T_{(3)j}, C_j)$ and $\{\Delta_{(r)j}, r = 1, 2, 3\}$ be the indicator function, which is one if $T_{(r)j} \leq C_j$. The data consist of $\{(X_{(1)ij}, X_{(2)ij}, X_{(3)ij}, \Delta_{(1)ij}, \Delta_{(2)ij}, \Delta_{(3)ij}, \delta_{ij}, U_{ij})', i = 1, \cdots, n_j\}$, n_j independent copies of $\{(X_{(1)j}, X_{(2)j}, X_{(3)j}, \Delta_{(1)j}, \Delta_{(2)j}, \Delta_{(3)j}, \delta_j, U_j)', j = 1, 2\}$.

Now, suppose that we are interested in estimating the t_0-year outcome probabilities $\pi_{jk}(U), j = 1, 2$, where

$$\pi_{jk}(U) = \text{pr}(\epsilon_j = k|U) \tag{2.1}$$

for a pre-specified time point t_0. To obtain estimates for $\pi_{jk}(U)$, one may use an ordinal regression working model of the following form

$$g(\gamma_{jk}(U_{ij})) = \alpha_k - \beta' Z_{ij} - \tau_{ij}(\beta^{*'} Z_{ij}) \tag{2.2}$$

where $\gamma_{jk} = \sum_{l=0}^{k} \pi_{jl}$ is the cumulative probability of a patient in treatment group j being classified into outcome category $\epsilon \leq k$, Z_{ij} is a given function of U_{ij}, $g(\cdot)$ is a given monotone function, and α, β_j and β_j^* are unknown vectors of parameters, with β_j and β_j^* corresponding to the main effects and treatment interaction effects, respectively, of the covariates Z on patient outcome status. Noting that a patient's outcome status is observable only when $\min(T_{(3)}, t_0) \leq C$, the parameter vectors above may be estimated by applying inverse probability of censoring weights and maximizing the standard weighted multinomial log-likelihood function

$$\sum_{ij} \frac{w_{ij}}{\hat{G}_j(X_{(3)ij} \wedge t_0)} \{\sum_{k=0}^{4} I(\epsilon_{ij} = k) log(\pi_{jk}(U_{ij}))\}, \tag{2.3}$$

with respect to (α, β, β^*) where $\pi_{jk} = f(\alpha, \beta, \beta^*, Z_{ij}, \tau_{ij})$ via (2.2), $w_{ij} = I(X_{(3)ij} \leq t_0)\Delta_{ij} + I(X_{(3)ij} > t_0)$, $I(\cdot)$ is the indicator function, and $\hat{G}_j(\cdot)$ is the Kaplan-Meier estimator for $G_j(\cdot)$, the survival function of the censoring variable for the jth group. (Zheng et al., 2006; Uno et al., 2007a; Li et al., 2011).

Under some mild conditions, the resulting estimators $(\hat{\alpha}, \hat{\beta}, \hat{\beta}^*)$ converge to a finite constant vector as $n \to \infty$ even when the model (2.2) is not correctly specified (Uno et al., 2007a).

Note that one may repeatedly utilize (2.2) and (3.5) with various Z and $g(\cdot)$ via, for instance, a standard stepwise regression procedure with U, to obtain final estimates $\hat{\pi}_{jk}(U)$.

2.3.1 Evaluation of Working Models

Since many variable selection procedures and link functions can be considered as candidates for estimating $\pi_{jk}(\cdot), j = 1, 2$, it is important to formally evaluate their relative merits. To this end, we first note that the adequacy of such ordinal regression models for t_0-year outcomes can be quantified by the cross-validated log-likelihood, where a larger log-likelihood suggests a better model fit. We use a repeated random cross-validation procedure, in each iteration randomly dividing the entire training data set into two mutually exclusive subsets, $\mathcal{B}$ and $\mathcal{H}$, with the "model building set" $\mathcal{B}$ comprising approximately 80% of the full training set. For each model building set and for a given link function and variable selection procedure, we can construct a model, using only patients in $\mathcal{B}$ to estimate $\pi_{jk}(U)$, yielding predicted probabilities for the patients in the holdout set $\mathcal{H}$, with predictions given by $\hat{\pi}_{jk}(U_{ij})$. The cross-validated log-likelihood, adjusted for censoring, is

$$\sum_{(i,j)\in\mathcal{H}} \frac{w_{ij}}{\hat{G}_j(X_{(3)ij} \wedge t_0)} \{\sum_{k=0}^{4} I(\epsilon_{ij} = k) log(\hat{\pi}_{jk}(U_{ij})\}, \tag{2.4}$$

We may further examine the adequacy of a given model in predicting outcomes for a specific treatment group by summing only over patients in a particular treatment group j.

We repeatedly split the training data set M times. For each m, and for each modeling procedure, we obtain an estimate of the group-specific log-likelihood value denoted by $\widehat{\mathcal{L}}_j^{(m)}$, and the overall log-likelihood $\widehat{\mathcal{L}}^{(m)} = \widehat{\mathcal{L}}_1^{(m)} + \widehat{\mathcal{L}}_2^{(m)}$. Lastly, we average these estimates over $m = 1, ..., M$ to obtain a final estimates $\widehat{\mathcal{L}}, \widehat{\mathcal{L}}_1$, and $\widehat{\mathcal{L}}_2$. The model(s) which yield the largest cross-validated likelihood values may be considered for the construction of our final working models. For each selected modeling procedure, we then refit the entire training data set with the specific modeling procedure in order to construct the final score.

In the BEST trial, 2708 patients were assigned to receive either placebo or bucindolol (a beta-blocker), with an average followup time of 2 years. In order to select a time point

representing reasonably long-term outcomes without substantial censoring, we use $t_0 = 18$ months (1.5 years) for the remainder of our analysis, with $\hat{G}_j(t_0) \approx 0.80$ for each treatment group. Within the BEST training set, 123, 60, 86, 9, and 84 patients in the control group were classified as healthy, alive without worsening HF, alive with worsening HF, non-CV death, and CV death, respectively after 18 months of followup. The corresponding counts for the treatment group were 148, 74, 52, 13, and 68 patients, respectively. The number of censored patients were 94 and 89 in control and treatment groups, respectively. We used 16 clinically relevant covariates to fit the patient outcome data with various working models to estimate the probability of each outcome status at $t_0 = 18$ months. These baseline variables are: age, sex, left ventricular ejection fraction (LVEF), estimated glomerular filtration rate adjusted for body surface area (eGFR), systolic blood pressure (SBP), class of heart failure (Class III vs. Class IV), obesity (Body mass index (BMI) $>$ 30 vs. BMI $\leq$ 30), resting heart rate, smoking status (ever vs. never), history of hypertension, history of diabetes, ischemic heart failure etiology, presence of atrial fibrillation at baseline, and race (white vs. non-white). As in Castagno et al. (2010), we used 3 indicator variables to discretize eGFR values into 4 categories, with cut-points of 45, 60, and 75. We used the ordinal regression models described above to fit the outcome data, and considered both the logit and complementary log-log links, $g(p) = log(\frac{p}{1-p}), g(p) = log(-log(1-p))$, respectively. For each of these two models, we used two different methods of variable selection. The first one used all 16 variables additively as well as their interaction terms. The second one used a stepwise regression procedure. It started from the full model including all 32 covariates (16×2) as well as the treatment indicator τ and successively added/eliminated terms until no more covariates could be added/removed without subsequently increasing the Akaike information criterion (AIC). Main effect terms were not eligible to be removed from the model unless the corresponding interaction term had already been removed. Therefore, a total of four modeling procedures were considered. To evaluate these models, we used a repeated random cross validation procedure as described above. In Table 2.3.1, we present these potential modeling procedures with their corresponding average log-likelihood values. It is interesting to note that most of these modeling procedures produce similar log-likelihood scores, but that the models using the complementary log-log link were generally found to produce more

accurate predictions in each treatment group. The full model using the complementary log-log link was found to provide the best overall fit in the cross-validation samples and will be used derive our scores in the following section. As a form of sensitivity analysis, we will also consider scores derived from the AIC-based model, also using the complementary log-log link, as this model was found to provide the best fit for the control group patients. We will refer to these models as the full model and AIC model, respectively.

Table 2.1: Candidate modeling procedures with average cross-validated log-likehoods

Link Function	Variable Sel.	$\widehat{\mathcal{L}}_1$	$\widehat{\mathcal{L}}$
logit	Full	-108.29	-202.30
logit	AIC	-107.95	-201.25
cloglog	Full	-107.56	-201.18**
cloglog	AIC	-107.52**	-201.29

Table 2.2: Ordinal regression coefficients for the final working models using BEST training data

	Full Model		AIC Model	
Covariate	β	β^*	β	β^*
Age	-0.001	-0.004	-	-
Sex: male	0.098	-0.148	-	-
LVEF	-0.014	-0.019	-0.014	-0.020
I(eGFR>75)	-0.175	-0.266	-0.191	-0.278
I(eGFR>60)	-0.041	-0.106	-	-
I(eGFR>45)	-0.656	-0.078	-0.733	-
SBP	-0.012	0.008	-0.011	0.007
Class IV Heart Failure	0.191	0.634	0.228	0.601
I(BMI>30)	0.207	0.010	0.197	-
Never-smoker	0.126	-0.128	-	-
Heart Rate	0.005	-0.015	0.006	-0.015
History of hypertension	0.213	-0.173	0.147	-
History of diabetes	0.308	-0.233	0.171	-
Ischemic etiology	0.083	0.103	-	-
Atrial Fibrillation	0.244	-0.286	-	-
Race: white	0.044	-0.198	-	-
Treatment	-	1.268	-	0.715

β represents main effects in fitted model. β^* represents treatment interaction effects.

2.4 Potential Scoring Systems

Having completed this variable selection and model building step to select the "best" working models for predicting patient outcomes in both treatment groups, there are two reasonable ways in which to incorporate covariate information for the purposes of stratifying patients in the BEST trial. Let treatment group $j = 1$ denote the untreated (placebo) group, and $j = 2$ denote the treated (beta-blocker) group.

2.4.1 Baseline Risk Score

Perhaps the most commonly used method for stratifying patients is by estimated baseline risk, indexed here by $\beta' Z_{ij}$. Denote this score as $\hat{R}(U)$. If all modeling assumptions made in the training data are valid, this linear combination of baseline covariates can be transformed to give consistent estimates for $\{\pi_{1k}(U), k = 0, \ldots, 4\}$, a patient's true outcome probabilities if not assigned to treatment. Even if the modeling assumptions are not correct, it is still plausible to assume that patients with larger risk scores may be generally at higher risk for more severe clinical outcomes than those with lower risk scores. A recent paper addressing the appropriate treatment of heart failure patients is indicative of the common clinical focus on high-risk vs. low-risk patient status, as well as the assumption that absolute treatment benefits are greatest for high-risk patients (Peterson et al., 2010).

2.4.2 Treatment Selection Score

Another interesting, though perhaps less commonly used, method for stratifying patients is according to treatment selection score (TSS), which is indexed by $\beta^{*'} Z_{ij}$, the model-based estimate of the differential effect of treatment for a patient with given covariates. For a given U, let this score for the treatment contrast be denoted by $\hat{D}(U)$, which intends to estimate

$D(U) = g(\gamma_{1k}(U)) - g(\gamma_{2k}(U))$ for any outcome category k if the modeling assumptions are true. Since γ_{jk} refers to the probability of being in a category equal to, *or healthier than*, category k, negative values of $D(U)$ correspond to reduction in overall risk associated with treatment. If patient i has true values $\gamma_{i1k} < \gamma_{i2k}, \forall k$, then that patient would unquestionably benefit from treatment. Conversely, if that patient's $\gamma_{i1k} > \gamma_{i2k}, \forall k$, then that patient would be universally harmed by treatment. While perhaps less intuitive for clinical use, it is clear that if the modeling assumptions are valid, then this score directly addresses the question of whether or not a particular patient is a good candidate for treatment. This type of stratification system has recently been implemented successfully by Cai et al. (2011), who investigated CD4 changes in HIV patients.

2.5 Making Inferences About the Treatment Differences over a Range of Scores with Respect to Ordered Patient Outcomes in the Target Data Set

Let the final parametric score for a patient with the covariate vector U in the target study be denoted by S(U), which may be the risk score $\hat{R}(U)$, based on the control group only, or the treatment selection score $\hat{D}(U)$ discussed in the previous section.

In order to make inference about the risks and benefits of treatment at the patient level, we propose to use the same ordered multinomial classification scheme described previously, in which each patient is classified according to their status at time t_0, depending on the particular set of clinical events which they have experienced. We then construct the confidence interval and band estimates for the treatment differences with respect to the probability of a patient being classified into each possible clinical category.

The target data consist of n_j independent and identically distributed observations as described in the training data. Furthermore, we let $\{Y_{ijk} = I(\epsilon_{ij} = k),\ k = 0, \ldots, 4\}$, which

is observable only if $\min(T_{(3)}, t_0) \leq C$. For the kth outcome, we are interested in estimating the treatment difference conditional on $S(U) = s$, that is,

$$E_k(s) = \text{pr}(\epsilon_{i2} = k|S(U) = s) - \text{pr}(\epsilon_{i1} = k|S(U) = s). \tag{2.5}$$

To estimate $E_k(s)$ nonparametrically, we use a kernel estimator for each term on the right hand side of Equation (2.5). Specifically, we estimate $p_{jk}(s) = \text{pr}(\epsilon_{ij} = k|S(U) = s)$ with $\hat{p}_{jk}(s)$

$$= \left\{ \sum_i^{n_j} \frac{w_{ij}}{\hat{G}_j(X_{(3)ij} \wedge t_0)} K_{h_j}(V_{ij} - s) Y_{ijk} \right\} / \left\{ \sum_i^{n_j} \frac{w_{ij}}{\hat{G}_j(X_{(3)ij} \wedge t_0)} K_{h_j}(V_{ij} - s) \right\}, \tag{2.6}$$

where $V_{ij} = S(U_{ij})$, $w_{ij} = I(X_{(3)ij} \leq t_0)\Delta_{(3)ij} + I(X_{(3)ij} > t_0)$, $\hat{G}_j(\cdot)$ is the Kaplan-Meier estimator of $G_j(\cdot)$, the survival distribution of the censoring variable C_j, estimated using observations $\{(X_{(3)ij}, \Delta_{(3)ij}), i = 1, \cdots, n_j\}$, $K_{h_j}(s) = K(s/h_j)/h_j$, $K(\cdot)$ is a smooth symmetric kernel with finite support and h_j is a smoothing parameter. Even though Y_{ijk} may not be observable due to censoring, note that $w_{ij}Y_{ijk}$ is always observable. When $h_j = O(n^{-v}), 1/5 < v < 1/2$, it follows from a similar argument by Li et al. (2011) that $\hat{p}_{jk}(s)$ converges to $p_{jk}(s)$ uniformly over the interval $s \in \mathcal{S}$, where $\mathcal{S}$ is an interval contained properly in the support of $S(U)$. Let $\mathbf{E}(s) = \{E_0(s), ..., E_4(s)\}' = \mathbf{p}_2(s) - \mathbf{p}_1(s)$ and its empirical counterpart $\hat{\mathbf{E}}(s) = \{\hat{E}_0(s), ..., \hat{E}_4(s)\}' = \hat{\mathbf{p}}_2(s) - \hat{\mathbf{p}}_1(s)$, where $\hat{E}_k(s) = \hat{p}_{2k}(s) - \hat{p}_{1k}(s)$, $\mathbf{p}_j(s) = \{p_{j0}(s), \cdots, p_{j4}(s)\}'$ and $\hat{\mathbf{p}}_j(s) = \{\hat{p}_{j0}(s), \cdots, \hat{p}_{j4}(s)\}'$

It follows from a similar argument by Li et al. (2011) that when h_j is of the same order as above, for a fixed s, the joint distribution

$$(n_1h_1 + n_2h_2)^{1/2}\{\hat{\mathbf{E}}(s) - \mathbf{E}(s)\} \tag{2.7}$$

converges in distribution to a multivariate normal with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}(s)$ as $n \to \infty$.

To approximate the distribution in (2.7), we use a perturbation-resampling method, which is similar to 'wild bootstrapping' (Wu, 1986; Mammen, 1993) and has been successfully implemented in many estimation problems (Lin et al., 1993; Park and Wei, 2003a; Cai et al.,

2010). Details are provided in the Appendix. To construct a $(1-\alpha)$ simultaneous confidence band for $E_k(s)$ over the pre-specified interval $\mathcal{S}$, we also use resampling methods, further described in the Appendix.

As with any nonparametric estimation problem, it is important that we choose appropriate smoothing parameters in order to make inference about $\mathbf{E}(s)$. Here, we use an M-fold cross-validation procedure to choose the smoothing parameter $\hat{h}_j$ which maximizes a weighted cross-validated multinomial log-likelihood, as in Li et al. (2011). Specifically, we may randomly divide the entire data set into M mutually exclusive, approximately equally sized subsets. For any fixed values of h_j and (j,k), we can estimate $p_{jk}(s)$ using all observations except for those contained in the same subset as the i^{th} subject, which yields the estimator $\hat{p}_{(-i)jk}(s)$. The cross-validated log-likelihood, adjusted for censoring, is

$$\sum_{V_{ij}\in\mathcal{S}} \frac{w_{ij}}{\hat{G}_j(X_{(3)ij} \wedge t_0)} \{\sum_{k=0}^{4} Y_{ijk} log(\hat{p}_{(-i)jk}(V_{ij}))\}. \tag{2.8}$$

Let $\hat{h}_j$ be a maximizer of (2.8). As in Li et al. (2011), $\hat{h}_j$ is of the order $n^{-1/5}$. To ensure the the bias of the estimator is asymptotically negligible and that the above large-sample approximation is valid, however, we slightly undersmooth the data and let the final smoothing parameter be $\tilde{h}_j = \hat{h}_j \times n^{-\xi}$ where ξ is a small positive number less than 0.3.

In order to aid in the interpretation of patient outcome probabilities, we may additionally estimate patient-specific cumulative probabilities by repeating the same procedure as above, but instead using $\tilde{Y}_{ijk} = I(\epsilon_{ij} \leq k), \gamma_{jk}(s) = E(\tilde{Y}_{ijk}|s)$, and $\Gamma_k(s) = \gamma_{2k}(s) - \gamma_{1k}(s)$. It should be noted that each value $\Gamma_k(s)$ allows for the estimation of the treatment contrast with respect to a different composite outcome. For example, $\Gamma_0(s)$ refers to the effect of treatment on the composite outcome "any hospitalization or death"; $\Gamma_1(s)$ refers to the effect of treatment on the composite of "CV hospitalization or death"; $\Gamma_2(s)$ corresponds to the effect of treatment on "any death", thus representing the initial primary outcome in the BEST study; $\Gamma_3(s)$ corresponds to the effect of treatment on CV-related death; $\Gamma_4(\cdot) \equiv 0$ by definition, as $\gamma_{2,4}(\cdot) = \gamma_{1,4}(\cdot) \equiv 1$. While a positive (negative) value of particular component

$E_k(s)$ may not always directly indicate whether a treatment is beneficial (harmful) for a specific patient, particularly when $0 < k < 4$, the corresponding $\Gamma_k(s)$ will always have this desired interpretation, with positive values always indicating beneficial treatment effects, as our classification scheme orders patient outcomes according to increasing severity.

2.5.1 Patient Outcomes in BEST Trial

First, we present for each treatment group, the total number of patients in the target data set known to be in each outcome category, as well as the estimated cell probabilities after adjusting for censoring. These results are shown in Table 2.5.1. We note that, overall, treated patients are somewhat more likely to be classified into outcome categories 0 and 1 (alive with no hospitalizations or non-CV hospitalizations only), and less likely to be classified into categories 2 and 4 (alive with CV-related hospitalization and CV death, respectively). The cumulative probabilities of a treated patient being classified at or below a certain threshold suggest an estimated population-level beneficial effect of treatment, regardless of the threshold used.

Table 2.3: BEST target data, 18-month patient outcomes: Observed patient outcomes and associated multinomial probability estimates, adjusted for censoring.

	Control Group			**Treated Group**		
Outcome Category	N	$P(\epsilon = k)$	$P(\epsilon \leq k)$	N	$P(\epsilon = k)$	$P(\epsilon \leq k)$
(censored)	156	-	-	159	-	-
0	274	0.384	0.384	294	0.405	0.405
1	114	0.160	0.544	150	0.207	0.612
2	165	0.231	0.775	138	0.190	0.802
3	26	0.031	0.806	26	0.031	0.833
4	162	0.194	1.000	143	0.167	1.000

Now, we apply the final scoring systems to the patients in the BEST trial. In Figure 2.5.1 below, we show the two scores $\hat{R}$ and $\hat{D}$ for each patient in the BEST target data set, as derived from the full model described previously, with parameters estimated using the training set data. It is interesting to note that the two sets of scores show no significant correlation. The horizontal line indicates $\hat{D} = 0$, and we note that 1284 (71%) of the target

BEST patients fall below this line, indicating an anticipated treatment benefit for a majority of patients.

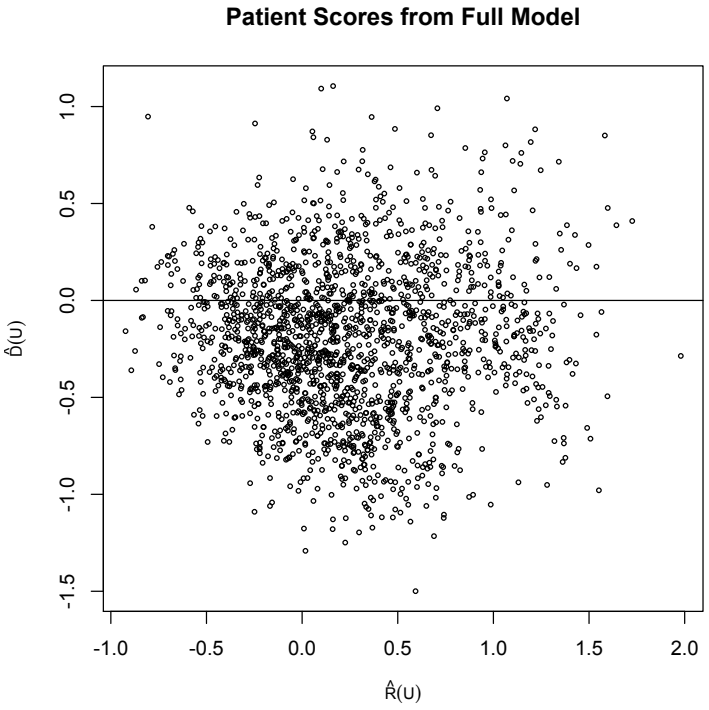


Figure 2.1: Risk Scores (R) and Treatment Selection Scores (D) for target BEST patients using models derived from BEST training data.

To estimate $\mathbf{p}_1(s)$, $\mathbf{p}_2(s)$ and $\mathbf{E}(s)$ in our analysis below, we let $K(\cdot)$ be the standard Epanechnikov kernel. The smoothing parameters were chosen as the maximizers of (2.8), then multiplied by $n_j^{-0.05}$.

Risk Score:

We may first validate our model-based risk score by observing the relative frequency of patient outcomes in the BEST placebo group only, as a function of the baseline risk score $\hat{R}$. In the left panel of Figure 2.5.1 below, the estimated control and treatment group outcome probabilities (solid and dashed lines, respectively) $\hat{\mathbf{p}}(s)$ are displayed over the range

$s \in (-0.70, 1.40)$. We note that the parametric risk score $\hat{R}$ is closely related to the non-parametrically estimated probabilities $\hat{p}_0(s)$ and $\hat{p}_4(s)$. That is, increasing risk scores are associated with increased risk of experiencing CV death ($\epsilon = 4$), and decreasing risk of remaining completely event-free ($\epsilon = 0$) in both treatment groups. The other three outcomes ($\epsilon = 1, 2, 3$) show relatively little overall association with the risk score. In the right panel, we see the estimated cumulative probabilities, representing the probability of a patient being classified at, or lower than, a given threshold. As expected, we see that, regardless of the threshold used, a patient's risk declines with increasing risk scores. In the bottom right panel, we present the distribution of the risk scores used in this analysis.

In order to determine whether the baseline risk score can be effectively used for patient-level treatment decisions, we evaluate the estimated treatment effects $\hat{\mathbf{E}}(s)$ and $\hat{\mathbf{\Gamma}}(s)$ over the range of scores. In the right panel of Figure 2.5.1, we show the treatment effects, reflecting the difference in outcome probabilities (treated minus untreated) for each patient outcome category. Estimated 95% confidence intervals are denoted by dashed lines, and 95% confidence bands are represented by the shaded areas. In general, we find that lower-risk patients (e.g. $\hat{R} < 0$), are generally more likely to be classified into outcome category 1 (non-CV hospitalization only), and less likely to be classified into outcome categories 2 (CV hospitalization) and 4 (CV death) as a result of treatment, while high-risk patients (e.g. $\hat{R} > 1$) are more likely to be classified into outcome category 0 (no clinical events) as a result of treatment, though the variability of the estimates increases in this range. The only treatment effects found to be significant with respect to the 95% confidence bands are associated with lower risk scores, as scores in the range (-0.33, 0.00) are associated with significant increases in the probability of experiencing the outcome $\epsilon \leq 1$ (alive without CV hospitalization) as a result of treatment. The more restrictive range (-0.24, -0.12) is further associated with significant increases in experiencing the outcome $\epsilon \leq 2$ (alive) and $\epsilon \leq 3$ (no CV death) due to treatment. These ranges represent approximately 23% and 8% of the patients in the BEST population, respectively.

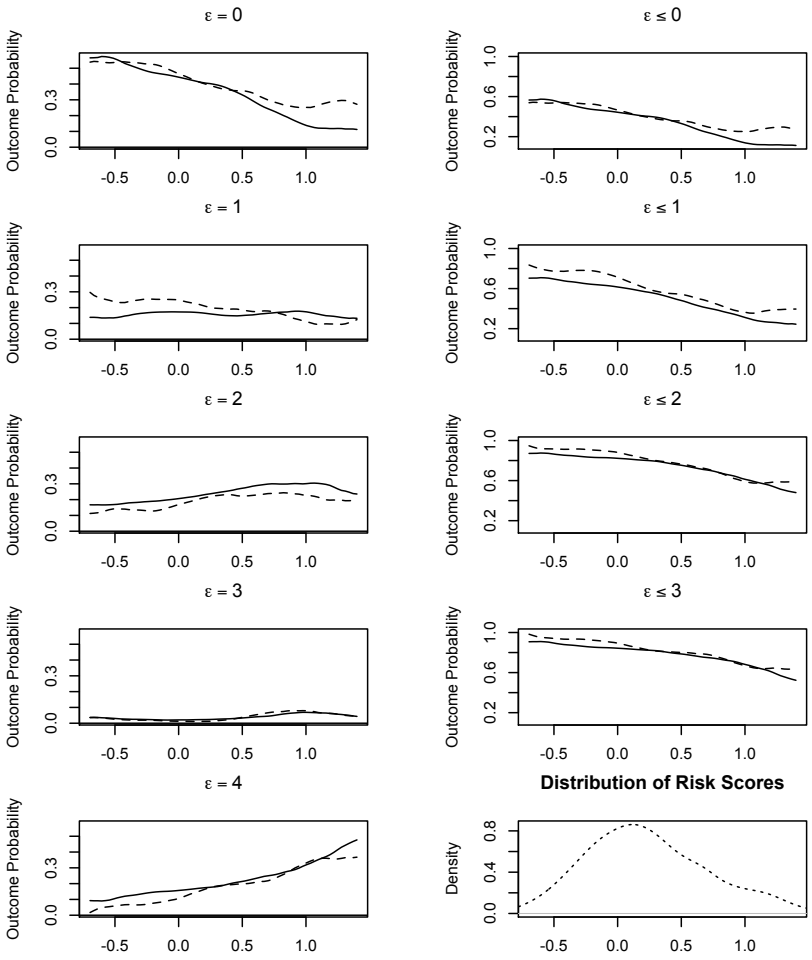


Figure 2.2: BEST target patient outcomes using baseline risk ($\hat{R}$) as a scoring system.
Solid lines: Placebo Group; Dashed lines: Treatment Group
Left panels: specific outcome probabilities; Right panels: cumulative outcome probabilities
Bottom right: Distribution of scores

Treatment Selection Score:

Using the TSS, we find that $\hat{D}$, a patient's model-based estimated effect of treatment on outcome status is related to the nonparametrically estimated treatment effect on a patient's probability of experiencing outcome category $\epsilon = 0$ (alive without any hospitalizations) and $\epsilon = 2$ (alive with CV hospitalization), $\hat{E}_0(s)$ and $\hat{E}_2(s)$. Specifically, using TSS as our scoring system over the range $s \in (-1.08, 0.76)$, we find $\hat{E}_0(s) > 0$ for $s < -0.08$ and $\hat{E}_0(s) < 0$ for $s > -0.08$, indicating that our treatment interactions estimated from the BEST training

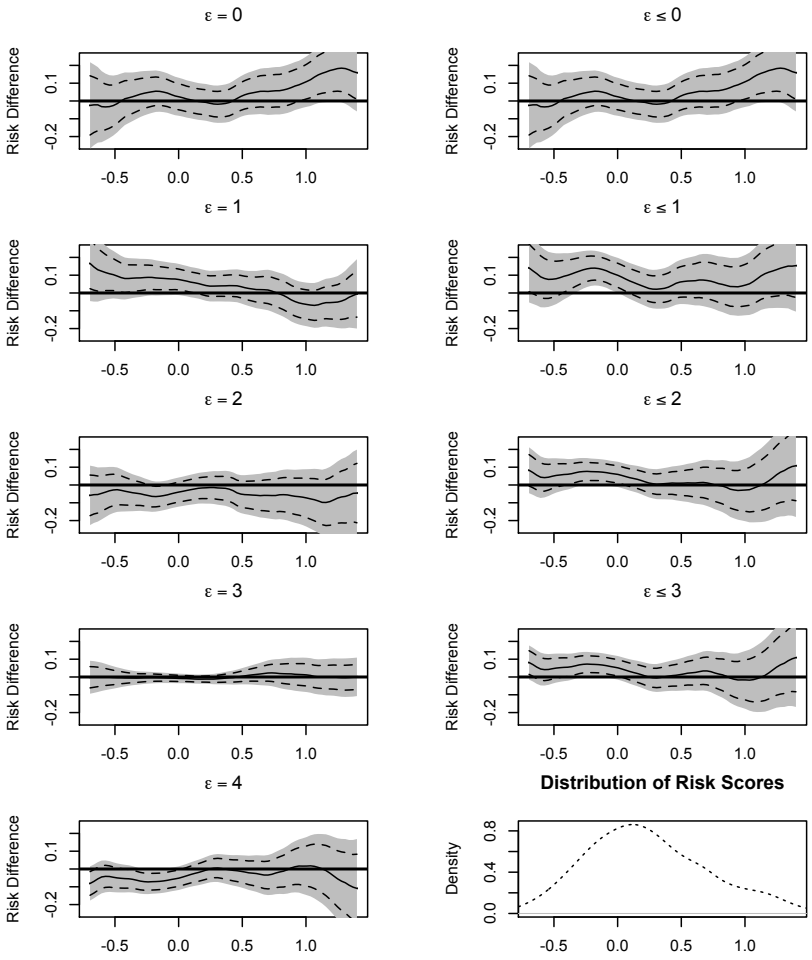


Figure 2.3: BEST target treatment differences (treated minus untreated) using baseline risk ($\hat{R}$) as a scoring system.
Left panels: specific outcome probabilities; Right panels: cumulative outcome probabilities
Bottom right: Distribution of scores

data appear to be helpful in predicting which patients would benefit from being treated with beta-blockers, though not necessarily with respect to mortality outcomes. In particular, patients with TSS $\in (-0.94, -0.30)$ are found to be significantly more likely (via the 95% confidence band) to experience outcomes $\epsilon \leq 1$ (alive with no CV hospitalization), a range of scores representing approximately 36% of the patients in the target data set. Treatment effects $\hat{\mathbf{E}}(\cdot)$ and $\hat{\mathbf{\Gamma}}(\cdot)$ are shown below in Figure 2.5.1. It is interesting to note that the estimated effect of treatment in terms of reducing the risk of death is relatively constant, and nonsignificant, with a risk reduction of approximately 2% across the range of scores.

Thus, it may be fair to declare a particular future patient to be a good or bad candidate for treatment on the basis on the treatment selection score, which is able to identify patients who would likely benefit from treatment in terms of avoiding CV-related hospitalization.

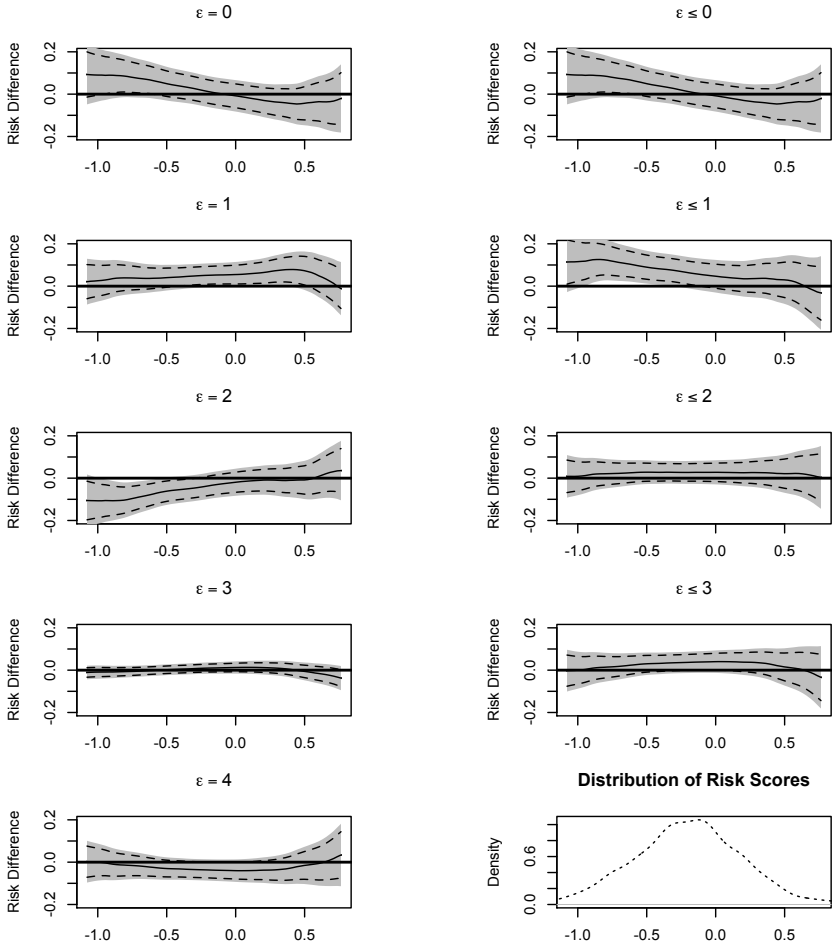


Figure 2.4: BEST target treatment differences (treated minus untreated) using treatment selection score ($\hat{D}$) as a scoring system.
Left panels: specific outcome probabilities; Right panels: cumulative outcome probabilities
Bottom right: Distribution of scores

2.5.2 Sensitivity Analysis using AIC model

Because our cross-validation procedure did not suggest that any particular model would fit the training data substantially better than the other candidate models, we repeat the above

procedure using the risk score and treatment selection score, as derived from the AIC model. From this model, there is a slight but significant positive correlation ($\rho = 0.17$) between the risk score and the treatment selection scores, suggesting that low-risk patients are generally expected to benefit more greatly from treatment than high-risk patients. In Figures 2.5.2 and 2.5.2 below, the treatment differences as a function of each of these scores are shown along with 95% confidence bands and intervals. The results are generally the same as when using the scores deriving from the full model. The most notable difference is that, using the treatment selection score deriving from the AIC model, in addition to detecting a range of patients who would be benefit significantly, via the 95% confidence intervals, in terms of avoiding CV-related hospitalization and/or death ($\epsilon \leq 1$), representing approximately 50% of the BEST population, a smaller subset of these patients, with TSS $\in (-0.64, -0.32)$ are found to benefit from treatment with respect to overall mortality and/or CV-related death ($\epsilon \leq 2, \epsilon \leq 3$). This smaller subset of scores is associated with approximately 25% of the BEST patient population.

2.5.3 Conclusions

Ultimately, despite the non-significant overall result in the BEST trial, both of our scoring systems are able to identify a sizeable subgroup of BEST patients who would experience significant benefits from treatment with beta-blockers (i.e. bucindolol), in each case representing approximately one fourth to one half of the BEST patient population, depending on whether 95% confidence intervals or bands are being used to determine statistical significance.

The results from our analysis using the baseline risk score are particularly interesting in that the observation that low-risk patients may derive equal (or possibly greater) benefits from treatment than high-risk patients seems to oppose clinicians' conventional wisdom. This finding may have implications for design of future studies.

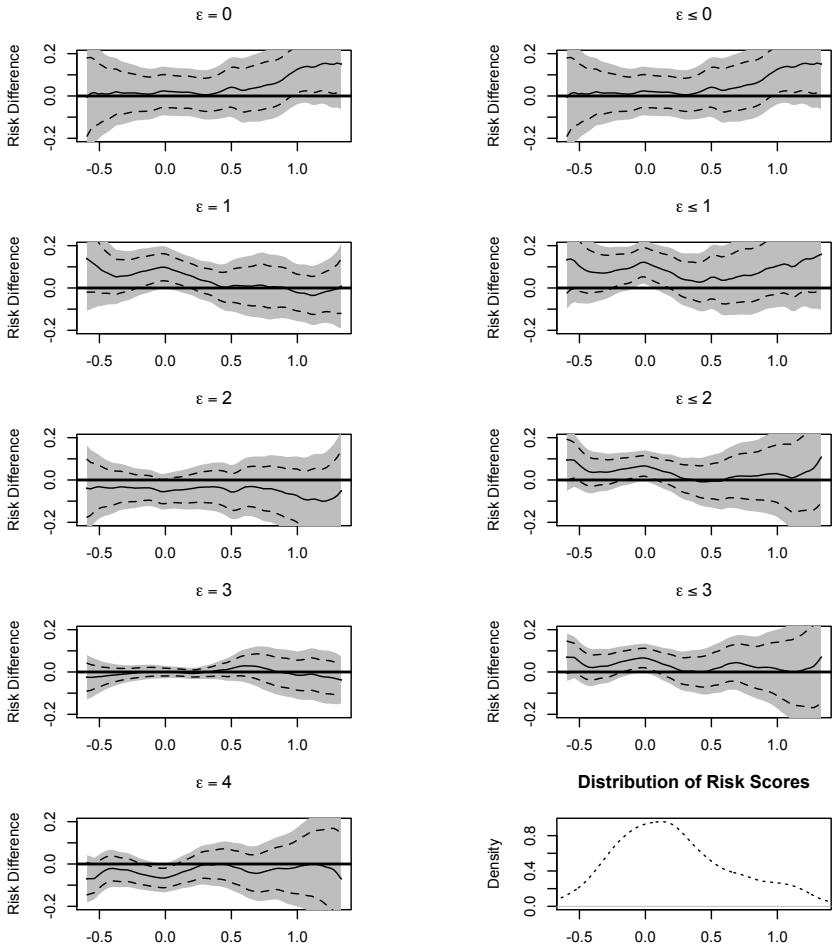


Figure 2.5: BEST target treatment differences (treated minus untreated) using baseline risk ($\hat{R}$) from the AIC model as a scoring system.
Left panels: specific outcome probabilities; Right panels: cumulative outcome probabilities
Bottom right: Distribution of scores

The results from our analysis using the treatment selection score are interesting for the reason that our scoring system, though not perfectly predictive of all clinical outcomes, seems to have done a reasonably good job of separating patients who would respond well to treatment (TSS<0) from those who would respond poorly (TSS>0). In this sense, we have found evidence of treatment interactions that are identifiable early in the course of the clinical trial and which are prospectively validated using future patients.

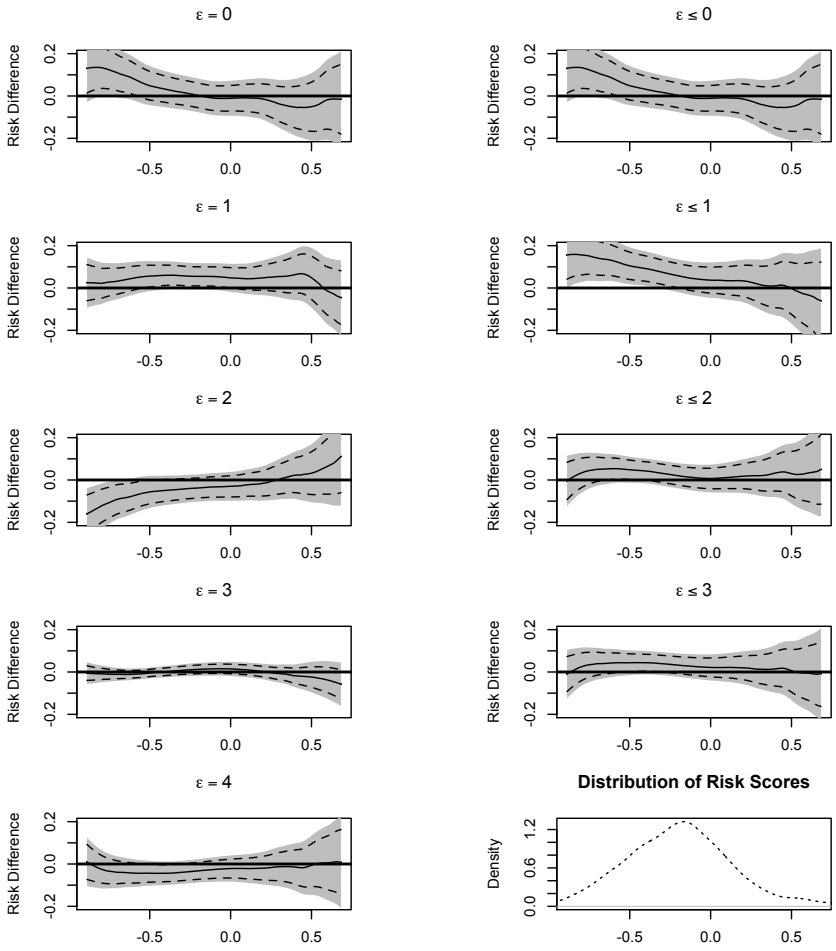


Figure 2.6: BEST target treatment differences (treated minus untreated) using treatment selection score ($\hat{D}$) from the AIC model as a scoring system.
Left panels: specific outcome probabilities; Right panels: cumulative outcome probabilities
Bottom right: Distribution of scores

2.6 Remarks

In this analysis, we use a two-stage process that can rather easily be applied in other scenarios with randomized clinical trial data. Here, we use two independent data sets obtained by separating the patients in a single clinical trial into two groups according to the order in which they were enrolled, thus utilizing a training data set similar to that which may be available at the time of an interim analysis. We utilize these independent data sets to construct a systematic, subject-specific treatment evaluation procedure. The final scoring

system may be chosen via a complex, exploratory model and variable selection process using the training data set. We then apply this system to stratify the patients in the second group and make inferences about the treatment effects with respect to various patient outcomes for each stratum. If two similar studies are available, as in many industry settings where two Phase III trials are required, one may instead treat the data from the first trial as the training data set and stratify and evaluate patients in the second trial accordingly.

Because our goal is to develop a score to group patients with similar treatment responses, the treatment selection score $\hat{D}(\cdot)$ should be the most effective system with which to stratify patients. However, there are practical concerns that could result in the preference of scoring system $\hat{R}(\cdot)$ based on the control group only. In particular, if there are multiple treatments options to be compared, it will not be feasible to build a separate treatment selection scores for each pairwise treatment comparison, and the baseline risk score would be an intuitive choice to investigate the effects of each treatment. Additionally if there are two trials of interest in the same disease setting but the patient populations and/or treatments are not known to be comparable across trials, then the treatment selection score $\hat{D}$ may not generalize well from one trial to another. For example, trials involving HIV-infected patients continually evaluate different combinations of approved therapies. Moreover, clinical practitioners seem more comfortable at present with the idea of using a risk score from the control arm to make treatment decisions; we show, however, in Figure 2.5.1 that the common belief that the patients with the worst prognosis will benefit the most from treatment may not hold true.

In this paper, we used the t_0-year outcome probabilities as the parameters of interest, where t_0 may be chosen from a clinical perspective. In practice, one may repeat this procedure with various time points. It would be interesting to choose a global measure to quantify the treatment contrast. Further research is warranted along this line.

Our model and variable selection procedure is intended to select the "best" model for fitting the data in terms of overall log-likelihood. When the endpoint is the treatment

difference, it is not clear that our approach of attempting to find a working model which best fits the data in this general sense would necessarily produce the "best" treatment selection score, as interactions between treatment and covariates become more important, while main effect terms may be thought of as nuisance parameters. In his unpublished thesis, Signorovitch (2007) proposed a novel method for modeling the treatment contrast directly with covariates for binomial outcomes. Intuitively, this approach would more effectively select the treatment and covariate interactions for creating the score. Further research is needed to evaluate scoring systems with respect to the subject-specific treatment differences.

Lastly, the choice of treatment based on a risk-benefit perspective is quite individualized. A global summary of the treatment *effectiveness*, for example, the risk-benefit ratio, may not provide enough information for personalized medicine. Instead, summaries for the treatment's combined toxicity and efficacy at the subject level, as we have proposed in this article, can be quite useful for the patient's bedside management.

2.7 Appendix

Construction of Confidence Intervals and Bands

Let $\{B_{ij} : i = 1, ..., n_j; j = 1, 2\}$ be independent random samples from a strictly positive distribution with mean and variance equal to one. Let $p^*_{jk}(s)$ be the perturbed version of $\hat{p}_{jk}(s)$ with $p^*_{jk}(s)$

$$= \left\{ \sum_i \frac{B_{ij} w_{ij}}{\hat{G}^*{}_j(X_{(3)ij} \wedge t_0)} K_{h_j}(V_{ij} - s) Y_{ijk} \right\} / \left\{ \sum_i \frac{B_{ij} w_{ij}}{\hat{G}^*{}_j(X_{(3)ij} \wedge t_0)} K_{h_j}(V_{ij} - s) \right\}. \quad (2.9)$$

Here, $\hat{G}^*_j(\cdot)$ is the perturbed estimator for the survival function $G_j(\cdot)$

$$\hat{G}^*_j(t) = \exp\left[- \sum_{i=1}^{n_j} \int_0^t \frac{B_{ij} d\{I(C_{ij} \leq u \wedge X_{(3)ij})\}}{\sum_{l=1}^{n_j} B_{lj} I(X_{(3)lj} \geq u)} \right]. \quad (2.10)$$

Denote $\mathbf{E}^*(s) = \mathbf{p}_2^*(s) - \mathbf{p}_1^*(s)$, where $\mathbf{p}_j^*(s) = \{p_{j0}^*(s), \cdots, p_{j4}^*(s)\}'$. Using the arguments by Cai et al. (2010), the limiting distribution, conditional on the target data set, of

$$(n_1h_1 + n_2h_2)^{1/2}\{\mathbf{E}^*(s) - \hat{\mathbf{E}}(s)\}, \tag{2.11}$$

is multivariate normal with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}(s)$.

In order to obtain an approximation to $\mathbf{\Sigma}(s)$, we generate a large number of realizations of $\{B_{i1}, B_{i2}\}$ from a standard exponential distribution, and compute $\mathbf{E}^*(s)$ for each perturbation sample. The resulting sample covariance matrix based on those perturbed estimates $\mathbf{E}^*$, say, $\tilde{\mathbf{\Sigma}}(s)$, is a consistent estimator of $\mathbf{\Sigma}(s)$. A two-sided confidence interval for an individual risk difference $E_k(s)$ is then given by

$$\hat{E}_k(s) \pm z_{(1-\alpha/2)}(n_1h_1 + n_2h_2)^{-1/2}\tilde{\sigma}_k(s), \tag{2.12}$$

where $\tilde{\sigma}_k(s)$ is the kth diagonal element of $\tilde{\mathbf{\Sigma}}(s)$.

To construct a $(1-\alpha)$ simultaneous confidence band for $E_k(s)$ over the pre-specified interval $\mathcal{S}$, we cannot use the conventional method based on the sup-statistic,

$$\sup_{s\in\mathcal{S}} \tilde{\sigma}_k^{-1}(s)|(n_1h_1 + n_2h_2)^{1/2}\{\hat{E}_k(s) - E_k(s)\}| \tag{2.13}$$

due to the fact that as a process in s, $(n_1h_1 + n_2h_2)^{1/2}\{\hat{E}_k(s) - E_k(s)\}$ does not converge to a process. On the other hand, one may utilize the strong approximation argument given in Bickel and Rosenblatt (1973) to show that an appropriately transformed sup of $\hat{E}_k(s) - E_k(s)$ converges to a proper random variable. In practice, to construct a confidence band, we can first find a critical value b_α such that

$$\text{pr}(\sup_{s\in\mathcal{S}} |E_k^*(s) - \hat{E}_k(s)|/\{(n_1h_1 + n_2h_2)^{-1/2}\tilde{\sigma}_k(s)\} > b_\alpha) \approx \alpha. \tag{2.14}$$

Then the confidence band for $E_k(s) : s \in \mathcal{S}$ is given by

$$\hat{E}_k(s) \pm b_\alpha(n_1h_1 + n_2h_2)^{-1/2}\tilde{\sigma}_k(s). \tag{2.15}$$

Identical arguments are used for making inference with respect to $\Gamma_k(s)$.

Acknowledgements

This manuscript was prepared using BEST Research Materials obtained from the NHLBI Biologic Specimen and Data Repository Information Coordinating Center and does not necessarily reflect the opinions or views of the BEST investigators or the NHLBI.

Summary of Treatment Impact Using Multiple Patient Outcomes

Brian Claggett[1], Lu Tian[2], Sihai Dave Zhao[1], and Lee-Jen Wei[1]

[1]Department of Biostatistics
Harvard School of Public Health

[2]Department of Health Research and Policy
Stanford University School of Medicine

3.1 Introduction

In many randomized studies, multiple events and associated failure times may be recorded for each patient. For example, in a recent cardiology study, event times corresponding to hospital admissions, myocardial infarction (MI), heart transplant, and death were all recorded. In such trials, it may be desirable for the investigators to assess the overall impact of treatment on patient health, as defined by the absence or occurrence of many events, while acknowledging a natural ordering to the severity of the events (e.g. hospital admission in general is less severe than MI, which is less severe than death). In such a situation, each patient may be evaluated at a fixed time and assigned an ordinal outcome status that acts as a summary of all clinical events experienced by a patient from randomization until the time of evaluation. In this paper, we propose methods to analyze such data over time, drawing connections to existing survival and ordinal regression techniques, allowing one to obtain a global summary of treatment impact.

Even though multiple patient outcomes are often observed in a study, investigators must often choose a single primary outcome, with the corresponding between-group comparison serving as the primary analysis of study, and upon which a decision as to the success of the treatment may be based. In recent years, seeking to gain efficiency, many trials have used combinations of clinical outcomes as a composite response score (Friedman et al., 2010). Noted challenges with this approach include the concerns that component outcomes may or may not have equal clinical importance, and that a single component may "dominate" the composite outcome (Ferreira-González et al., 2007; Tomlinson and Detsky, 2010). Methods such as Q-TWiST have been proposed in the context of cancer trials where patients are observed to transition from one health state to another over time (i.e., toxicity, time without symptoms and toxicity, relapse), but the procedure may be difficult to interpret when more than three states are involved, and inference depends on the specification of subjective utility weights associated with each health state (Gelber et al., 1995, 1996). Many methods have been proposed, primarily for the purposes of risk-benefit assessment, which classify patients

according to their overall response at the end of the trial (Chuang-Stein et al., 1991; Boers et al., 2010). These methods also generally require subjective utility weights and/or may ignore a patient's clinical experience over time. Follmann (2002) proposed the usage of multiple events for each patient in order to describe patients' overall clinical response, but did so for the purposes of ranking patients with respect to one another at the end of the trial. In this paper, we propose methods that estimate an overall summary of the effects of treatment and/or assess the effect of treatment as a function of time, taking into account patient health status as it changes over time, without the need to specify utility weights associated with each health state.

3.1.1 Notation: Progressive Disease State

Suppose we have an integer-valued variable denoting patient health status over time, $E(t) \in \{0, \ldots, K\}$, where larger values generally denote worse health. We focus our attention on the setting of a progressive disease model, defined by the following properties, and will discuss later when such assumptions may be relaxed. Suppose $E(t)$ has the following properties:

- a) $E(0) = 0$;
- b) $E(t) \leq E(s)\ \forall s > t$.

Now, for each patient, define the 'exit time' $T_k = min\{t : E(t) \geq k\}$,for $k \in \{1, \ldots, K\}$. Conversely, $E_i(t) = \sum_{k=1}^{K} I\{T_{ik} \leq t\}$.

Alternatively, suppose there are K types of "events" which a patient may experience, ordered by severity where event type K is the worst. For the ith patient $\{i = 1, \ldots, n\}$, define T^*_{ik} be the time to first instance of an event associated with severity level k. Let C_i be the patient's censoring time, which is assumed to be independent of T^*_{ik}. The observable data for each patient is $\{(X_{ik}, \Delta_{ik}), k = 1, \ldots, K\}$ where $X_{ik} = min(T_{ik}, C_i)$, $\Delta_{ik} = I(T_{ik} \leq C_i)$,

$T_{ik} = min\{T^*_{ik}, \ldots, T^*_{iK}\}$, and $I(\cdot)$ is the indicator function. We note that the vectors X_k and T_k will be ordered monotonically (i.e., $X_{i1} \leq X_{i2} \leq \ldots \leq X_{iK}$, $T_{i1} \leq T_{i2} \leq \ldots \leq T_{iK}, i = 1, \ldots, n$), even though the original event times T^*_k may not be ordered. The "event" associated with level k may be a singular event, a particular recurrence of an event, or perhaps a combination of events (e.g. death due to a particular cause). As such, some events may compete with other events, and an observed finite value of one T^*_k may necessarily result in a censored value for a competing T^*_j. Let $S_k(t) = P(T_k > t) = P(E(t) < k)$. Let Z denote treatment assignment ($Z = 1, 0$ for treated and untreated groups, respectively). The goal is to estimate the joint effect of treatment on the functions $\{S_k(t), k = 1, \ldots, K\}$. The relationship between treatment assignment and time-dependent patient outcome status may be analyzed via several different methods, some requiring stronger parametric assumptions than others. In this paper, we present three general approaches to this problem, with each subsequent approach requiring stronger assumptions than the last.

3.2 Integrated General Risk Difference

The general risk difference has been studied extensively as an extension of the simple risk difference for ordinal data with more than two outcomes (Simonoff et al., 1986; Edwardes and Baltzan, 2000; Lui, 2002). With two exposure categories, here $Z = 0, 1$, the general risk difference is $P(E_i > E_j | Z_i = 0, Z_j = 1) - P(E_i < E_j | Z_i = 0, Z_j = 1)$, with positive values suggesting that treated patients are generally more likely to be in a "healthier" state than their untreated counterparts. Evaluated at a fixed time t, this type of measure is closely related to Somers' d statistic as well as Wilcoxon's rank-sum statistic, and can further be shown to reduce to the simple risk difference $P(E = 1|Z = 0) - P(E = 1|Z = 1)$ when E is a binary response (Agresti, 1990; Edwardes, 1995). Specifically, define the general risk difference with respect to patient outcomes at time t as $GRD(t) = \sum_{k=1}^{K} \pi_{0,k}(t)\{\sum_{j=0}^{k-1} \pi_{1,j}(t)\} - \sum_{k=1}^{K} \pi_{1,k}(t)\{\sum_{j=0}^{k-1} \pi_{0,j}(t)\}$, where $\pi(t)_{j,k} = P(E(t) = k | Z = j)$. Let $\widehat{GRD}(t) = \sum_{k=1}^{K} \hat{\pi}_{0,k}(t)\{\sum_{j=0}^{k-1} \hat{\pi}_{1,j}(t)\} - \sum_{k=1}^{K} \hat{\pi}_{1,k}(t)\{\sum_{j=0}^{k-1} \hat{\pi}_{0,j}(t)\}$ where $\hat{\pi}_{j,k}(t)$ is a

consistent estimate of $\pi_{j,k}(t)$.

With complete data, the empirical estimates $\hat{\pi}(t)_{j,k} = \frac{\sum_i I(E_i(t)=k|Z=j)}{\sum_i I(Z_i=j)}$ could be used. Censoring will in general prevent the observation of $\{I(E_i(t) = k), t \geq C_i\}$ for any outcome level $k < K$. However, the nature of the progressive disease model implies that $I(E_i(t) \geq k)$ may be observable for some levels k. Specifically, $I(E_i(t) \geq k) = 1, \{k \geq E_i(C_i), t \geq C_i\}$. Noting that each group-specific survival curve $S_{j,k}(t) = P(E(t) < k|Z = j)$, we may define $\hat{\pi}_{j,k}(t) = \hat{S}_{j,k+1}(t) - \hat{S}_{j,k}(t)$, with $\hat{S}_{j,0}(\cdot) = 0, \hat{S}_{j,K+1}(\cdot) = 1$ for mathematical convenience.

Noting that the integrated difference in Kaplan-Meier curves is a commonly used model-free estimate of treatment effect, which is interpreted as "excess restricted mean survival time due to treatment," we introduce a generalization to this approach by integrating the estimate $\widehat{GRD}(t)$ over time. Define $IGRD(t) = \int_{s=0}^{t} GRD(s)ds$, and its estimator $\widehat{IGRD}(t) = \int_{s=0}^{t} \widehat{GRD}(s)ds = \int_{s=0}^{t} \hat{P}(E_i(s) > E_j(s)|Z_i = 0, Z_j = 1)ds - \int_{s=0}^{t} \hat{P}(E_i(s) < E_j(s)|Z_i = 0, Z_j = 1)ds$. The first component $\int_{s=0}^{t} \hat{P}(E_i(s) > E_j(s)|Z_i = 0, Z_j = 1)ds$ has the interpretation of average person-time spent with treated patients in better health than untreated patients. Conversely, $\int_{s=0}^{t} \hat{P}(E_i(s) < E_j(s)|Z_i = 0, Z_j = 1)ds$ has the interpretation of average person-time spent in which untreated patients are in better health than treated patients. Thus the difference of these two measures can be interpreted as "excess time spent in improved health due to treatment".

Alternatively, for any two patients (i, j) at some follow-up time t^*, the duration of followup may be partitioned into three components, such that $t^* = \int_0^t I(E_i(s) < E_j(s))ds + \int_0^t I(E_i(s) = E_j(s))ds + \int_0^t I(E_i(s) > E_j(s))ds = t_i^* + t_e^* + t_j^*$, representing the cumulative time in which patient i is in better, equal, and worse health than patient j, respectively. The difference in times $t_i^* - t_j^*$ represents the excess time spent in improved health for patient i relative to patient j. The value $IGRD(t)$ can be thought of as the expected value of this difference, conditioning on $Z_i = 1, Z_j = 0$. This alternative formulation is derived in the Appendix.

In order to approximate the variance of $\widehat{IGRD}(t) - IGRD(t)$, we may employ perturbation-resampling methods to estimate the variance of $\widehat{IGRD}^*(t) - \widehat{IGRD}(t)$, where

$$\widehat{IGRD}^*(t) = \int_{s=0}^{t} \widehat{GRD}^*(s)ds, \tag{3.1}$$

$$\widehat{GRD}^*(t) = \sum_{k=1}^{K} \pi_{0,k}^*(t)\{\sum_{j=0}^{k} \pi_{1,j}^*(t)\} - \sum_{k=1}^{K} \pi_{1,k}^*(t)\{\sum_{j=0}^{k} \pi_{0,j}^*(t)\},$$

$$\hat{\pi}_{j,k}^*(t) = \hat{S}_{j,k+1}^*(t) - \hat{S}_{j,k}^*(t),$$

and $\hat{S}_{j,k}^*(t)$ is the perturbed Kaplan-Meier estimator

$$\hat{S}_{j,k}^*(t) = \exp\left[- \sum_{i:Z_i=j} \int_0^t \frac{B_i d\{I(C_i \leq u \wedge X_{ik})\}}{\sum_{l:Z_l=j} B_l I(X_{ik} \geq u)} \right], \tag{3.2}$$

where each perturbed value $\widehat{IGRD}^*(t)$ employs a vector of standard exponential random variables $\{B_i, i = 1, \ldots, N\}$, generated independently of the data. This type of procedure has been applied successfully in previous work such as Park and Wei (2003b), Cai et al. (2005), and Uno et al. (2007b).

3.3 Repeated Ordinal Regression

Another option for quantifying the impact of treatment at a fixed follow-up time t_0 is via an ordinal regression model (McCullagh, 1980). In the generalization of ordinal logistic regression models known as the cumulative link model, the typical setup involves a categorical outcome variable Y with $\gamma_k = P(Y \leq k) = g^{-1}(\theta_k + X'\beta)$ for some general covariate vector X and a monotonic link function $g(\cdot)$ (Agresti, 1990). In our setup, the outcome is related to the treatment assignment via the model

$$g(\gamma_k) = \theta_k + Z'\beta, \tag{3.3}$$

If we let $Y_i = E_i(t_0)$ for a fixed time t_0, then each patient's log-likelihood contribution

is given by

$$\begin{aligned} L_i &= \prod_{k=0}^{K} \pi_{ik}^{I(Y_i=k)} \\ &= \prod_{k=0}^{K} \{\gamma_{ik} - \gamma_{i(k-1)}\}^{I(Y_i=k)} \\ &= \prod_{k=0}^{K} \{g^{-1}(\theta_k + Z_i'\beta) - g^{-1}(\theta_{k-1} + Z_i'\beta)\}^{I(Y_i=k)} \end{aligned} \tag{3.4}$$

With complete data, it is possible to maximize the above likelihood with respect with respect to $\{(\beta, \theta_k), k = 0, \ldots, K-1\}$, where $\gamma_{iK} \equiv 1, \gamma_{i(-1)} \equiv 0$ by definition. However, the independent censoring process will prevent the observation of patient outcomes if the patient was not in a "terminal state" at the time of censoring, as defined below.

3.3.1 Definition and Notation for Terminal States

It is possible that the definitions of the clinical events used to construct the patient outcome ranking system may result in a set of "competing" outcomes. For example, if the event "death due to heart failure (HF)" is associated with patient outcome level K, while "non-HF death" is associated with outcome level $K-1$, then it is clear that T_{K-1}^* will censor T_K^* for a given patient, and vice versa. Thus the observation of T_{K-1}^* indicates that a patient is in a terminal state, and, for such a patient, $P[E(t) > (K-1)] = 0$ for any time t. Specifically, we define a vector $\mathcal{V}$ indicating terminal states, where $\mathcal{V}_k = 1$ if $P(\sum_{j \geq k} \Delta_{ij} = 1 | \Delta_{ik} = 1) = 1$ and $\mathcal{V}_k = 0$ otherwise. In the progressive disease model, patients cannot exceed state K, and so $\mathcal{V}_K = 1$. Using this notation, let us further define for each patient $T_{iT} = min_k\{T_{ik}^* : \mathcal{V}_k = 1\}$, as the time to entering a terminal state, $X_{iT} = min(T_{iT}, C_i)$, and $\Delta_{iT} = I(T_{iT} \leq C_i)$. By the definition of a terminal state, $E_i(s)$ is known for any $s > T_{iT}$ when $\Delta_{iT} = 1$, and so each patient's status at a fixed t_0 is determined at $(t_0 \wedge T_{iT})$.

3.3.2 Weighted Cumulative Link Model

Using this information, we may modify the log-likelihood from the above cumulative link model by applying inverse probability of censoring weights and maximizing the standard weighted multinomial log-likelihood function

$$\sum_i \frac{w_i}{\hat{G}(X_{iT} \wedge t_0)} \{ \sum_{k=0}^{K} I(Y_i = k) log(\pi_{ik}) \}, \tag{3.5}$$

where $\pi_{ik} = g^{-1}(\theta_k + Z_i'\beta) - g^{-1}(\theta_{k-1} + Z_i'\beta)$, $\hat{G}(\cdot)$ is the Kaplan-Meier estimator of $G(\cdot)$, the survival function for the censoring variable C, and $w_i = I(C_i > t_0) + I(X_{iT} < C_i)\Delta_{iT}$ is one if a patient's outcome status is observable at time t_0 and zero otherwise. For a specific followup time t_0, we may denote the resulting maximizing value β as $\hat{\beta}(t_0)$.

In order to assess the treatment effect over time, one may repeat this procedure at multiple followup times $t_1 < t_2 < \ldots < t_J$. If we suspect that the true treatment effect may be constant over time (*i.e.*, $\beta(\cdot) = \beta$), then β may naturally be estimated via a linear combination $\sum_j c_j \hat{\beta}(t_j)$ with $\sum_j c_j = 1$. As in Wei and Johnson (1985) and Wei et al. (1989), the linear combination with smallest asymptotic variance is given by $c = (c_1, \ldots, c_J)' = (e'\hat{\Psi}^{-1}e)^{-1}\hat{\Psi}^{-1}e$ where $e = (1, \ldots, 1)$, and $\hat{\Psi}$ is the estimated covariance matrix of $(\hat{\beta}(t_1), \ldots, \hat{\beta}(t_J))$. The estimated variance for this linear combination is given by $(e'\hat{\Psi}^{-1}e)^{-1}$. If the treatment effect $\beta(\cdot)$ is constant in time, then such a linear combination will provide a consistent estimator for β, regardless of the followup times chosen. Even if the true value $\beta(\cdot)$ is not constant over time, this linear combination may still provide a useful summary measure for the "average effect" of treatment over time, however the expected value of this estimate will naturally depend on the specific times $\{t_j, j = 1, \ldots, J\}$ at which the model is fit (Wei et al., 1989). The covariance matrix $\hat{\Psi}$ may be estimated via a similar perturbation-resampling procedure to that described in the previous section. Specifically, for $r = 1, \ldots, R$, we generate a vector of exponential random variables $\{B_i^r\}, i = 1, \ldots, N$.

Then, for each time t_j, let $\hat{\beta}^{*r}(t_j)$ be maximizer of

$$\sum_i \frac{B_i^r w_i}{\hat{G}^*(X_{iT} \wedge t_j)} \{\sum_{k=0}^{K} I(Y_i(t_j) = k) log(\pi_{ik}(t_j))\}, \tag{3.6}$$

where $\hat{G}^*(\cdot)$ is the perturbed version of the Kaplan-Meier estimator $\hat{G}(\cdot)$, as in (3.2). Let $\hat{\beta}^*(t_j) = \{\hat{\beta}^{*r}(t_j)\}, r = 1, \ldots, R$. Then, each element $\hat{\Psi}_{j,j'}$ is estimated by $Cov(\hat{\beta}^*(t_j), \hat{\beta}^*(t_{j'}))$.

3.4 Global Model

Rather than repeatedly fitting models at multiple time points, we may instead wish to build the assumption of a constant treatment effect β directly into our model. Recall the cumulative link model described in the previous section. We have a categorical outcome variable Y with $\gamma_k = P(Y \leq k)$, and the outcome is related to covariates via the model

$$g(\gamma_k) = \theta_k + Z'\beta, \tag{3.7}$$

Letting $Y(t) = E(t)$, we can extend this model across time, writing

$$g(\gamma_k(t)) = \theta_k(t) + Z'\beta(t). \tag{3.8}$$

If we are willing to make the additional assumption that the association between treatment and outcome status $\beta(t)$ is constant for all t, then (3.8) reduces to

$$g(\gamma_k(t)) = \theta_k(t) + Z'\beta, \tag{3.9}$$

which we refer to as our proposed global model.

3.4.1 Relationship to semi-parametric survival models

Recall that, in general, when there is only a single event time, we may express many semi-parametric survival models as

$$g(S(t)) = \theta^*(t) + Z'\beta \tag{3.10}$$

for some decreasing function $g(\cdot)$ (Cheng et al., 1995). In our current framework, we have K such survival models, corresponding to event times $T_k, (k = 1, \ldots, K)$. Assuming the same link function $g(\cdot)$ for each model, we may write this collection of survival models as

$$g(S_k(t)) = \theta_k^*(t) + Z'\beta_k, \tag{3.11}$$

where $\theta_1^*(t) \geq \theta_2^*(t) \geq \ldots \geq \theta_K^*(t)$ for each fixed t. Finally, we note that

$$\gamma_k(t) = P(E(t) \leq k) = P(E(t) < k+1) = S_{k+1}(t). \tag{3.12}$$

Now, if we assume that $\beta_k = \beta \ \forall k$, we see that (3.11) reduces to (3.9), with nuisance parameters $\theta_k^*(t) \equiv \theta_{k-1}(t)$.

We note that, with a single binary covariate Z, a fully saturated model for the state probabilities over time can be written as

$$g(S_k(t)) = \theta_k(t) + Z'\beta_k(t), \tag{3.13}$$

where the treatment effect $\beta_k(t)$ is allowed to vary with time t and/or outcome status k.

The assumption that $\beta_k(t) = \beta_k \ \forall t$ results in the set of semi-parametric survival models (3.11). The assumption that $\beta_k(t) = \beta(t) \ \forall k$ results in the longitudinal cumulative link model (3.8). Our proposed model (3.9) results from the simultaneous assumption of these two relationships (i.e. $\beta_k(t) = \beta, \forall t, k$).

3.4.2 Estimating treatment effect with a single terminal state via stratified Cox model

When the vector $\mathcal{V} = \{0, \ldots, 0, 1\}$, (i.e., there is only a single terminal state), the estimation of the parameter of interest, β is relatively straightforward. Each set of exit times T_k can be thought of as a separate set of failure times to be modeled via a semi-parametric survival model, with associated score equation $U_k(\beta)$. In this case, we simply "stack" the data sets together, and sum the score equations to obtain the full score $U(\beta) = \sum_k U_k(\beta)$.

In the special case that the link function $g(\cdot) = log(-log(\cdot))$, then we note that (3.10) is equivalent to the Cox proportional hazards model, resulting in a partial likelihood $L(\beta) = \prod_{k=1}^{K} L_k(\beta)$, where

$$L_k(\beta) = \prod_{i=1}^{n} \left[\frac{exp\{Z_i'\beta\}}{\sum_{l \in \mathcal{R}_k(X_{ki})} exp\{Z_l'\beta\}} \right]^{\Delta_{ik}}, \tag{3.14}$$

where $\mathcal{R}_k(t) = \{l : X_{ik} \geq t\}$, which is similar to Wei et al. (1989) with the constraint that all $\beta_k = \beta$. We further note that the above likelihood takes the same form as a stratified Cox model with K strata, where the data corresponding to stratum k is simply $\{(X_{ik}, \Delta_{ik}, Z_i), i = 1, \ldots, n\}$. We denote the resulting maximizer of $L(\beta)$ as $\hat{\beta}$. Specifically, $\hat{\beta}$ is found as the solution to the score equation $U(\beta) = \sum_k U_k(\beta) = 0$, where $U_k(\beta) = \sum_{i=1}^{n} \Delta_{ik} \left\{ Z_i - \frac{\sum_i Y_{ik}(X_{ik}) Z_i exp\{Z_i'\beta\}}{\sum_i Y_{ik}(X_{ik}) exp\{Z_i'\beta\}} \right\}$, $Y_{ik}(t) = I(X_{ik} > t)$.

Because the vector of event times $\{X_{ik}\}$ are positively correlated within patients, the standard variance estimate $\tilde{V}(\hat{\beta})$ from the stratified Cox model will underestimate the true variance $V(\hat{\beta})$, and would result in an anti-conservative test of $H_0 : \beta = 0$. In the standard case, with only one terminal state, the robust variance estimator of Lin and Wei (1989) may be used. Here, $\hat{V}(\hat{\beta}) = \tilde{D}'\tilde{D}$, where $\tilde{D} = \tilde{U}\mathcal{I}^{-1}, \tilde{U}_i = \sum_k \tilde{u}_{ik}$, and $\tilde{u}_{ik}$ is the score residual corresponding to the kth event time from the ith patient, and $\mathcal{I} = \sum_k \mathcal{I}_k$, where $\mathcal{I}_k$ is the information matrix corresponding to the kth data set, $\{(X_{ik}, \Delta_{ik}, Z_i), i = 1, \ldots, n\}$. (Therneau and Grambsch, 2000). We may then assume $(\hat{\beta} - \beta) \sim N(0, \hat{V}(\hat{\beta}))$ and apply standard techniques for hypothesis testing and the construction of confidence intervals.

3.4.3 Extensions of the Global Model

Estimating Treatment Effect with Competing Risks

When there exist multiple terminal states (i.e. $\sum_k \mathcal{V}_k > 1$), we must introduce further notation to properly account for competing risks. For any $k : \sum_{j<k} \mathcal{V}_j = 0$, the score equation $U_k(\beta)$ will remain unchanged. For any $k : \sum_{j<k} \mathcal{V}_j > 0$, we must modify $U_k(\beta)$ as follows. First, we introduce additional notation. Specifically, let $\Delta_k^* = I(T_k^* \leq C)$ and $R_{ik} = \sum_{j<k} \mathcal{V}_j \Delta_{ij}^*$, which takes the value 1 if patient i is observed to enter into a terminal state which prevents the entrance into state k.

Using the methods of Fine and Gray (1999) methods for the subdistribution of a competing risk, the new score equation $U_k(\beta)$ is

$$\sum_{i=1}^{n} \Delta_{ik} \left\{ Z_i - \frac{\sum_i w_{ik}(X_{ik}) Y_{ik}^*(X_{ik}) Z_i exp\{Z_i'\beta\}}{\sum_i w_{ik}(X_{ik}) Y_{ik}^*(X_{ik}) exp\{Z_i'\beta\}} \right\} \tag{3.15}$$

where $Y_{ik}^*(t) = 1 - I(X_{ik} < t, R_{ik} = 0), w_{ik}(t) = I(t \wedge X_{ik} \leq C_i)\hat{G}(t \wedge T_{ik})/\hat{G}(t)$ and $\hat{G}(t)$ is the Kaplan-Meier estimate of the survival function of the censoring random variable C.

When there are multiple terminal states, it is not clear that robust variance procedures may be applied. For the purposes of hypothesis testing, a permutation test may be used, wherein the values Z_i may be permuted a large number of times R, resulting in vectors $\{Z_i\}_r, (r = 1, \ldots, R)$. For each permuted vector $\{Z_i\}_r$, the model is refit, and the resulting point estimate denoted $\hat{\beta}_r^*$. For a two-sided α-level test of the null hypothesis, we reject when $\frac{\sum_{r=1}^{R} |\hat{\beta}| \geq |\hat{\beta}_r^*|}{R} \leq \alpha$. Perturbation-resampling techniques and/or bootstrapping may be used to estimate a confidence interval for the parameter β.

Alternative Link Functions

For general link functions $g(\cdot)$, Cheng et al. (1995) show that (3.10) can be re-written as

$$h(T) = -Z'\beta + \epsilon, \tag{3.16}$$

where $\epsilon \sim F(\cdot) = 1 - g(\cdot)^{-1}$, and $h(t)$ is an unspecified monotone function increasing in t. They propose a score function based on the relationship between $P(T_i > T_j)$ and $(Z_i - Z_j)'\beta$ for all pairs (i, j). The resulting point estimate $\hat{\beta}$ is the solution to $U(\beta) = 0$, where $U(\beta)$ is given in eq. (2.3) in Cheng et al. (1995). This can be extended to current setup by solving for $U(\beta) = \sum_{k=1}^{K} U_k(\beta)$, where $U_k(\beta)$ is the stratum-specific score equation, equivalent to the score equation of Cheng et al. (1995) when using only $\{(X_{ik}, \Delta_{ik}, Z_i), i = 1, \ldots, n\}$. It is still possible to perform hypothesis testing and obtain confidence intervals via permutation tests and resampling techniques as mentioned above.

3.5 Example

Our data set of interest comes from a recent clinical trial, "Beta-Blocker Evaluation of Survival Trial" (BEST), which compared a beta-blocker (bucindolol) to placebo in patients with heart failure, with a primary endpoint of all-cause mortality. In this trial, other monitored patient outcomes included myocardial infarction (MI), all-cause hospitalization, and heart transplant (Beta-Blocker Evaluation of Survival Trial Investigators, 2001). Note that in the BEST trial, the non-fatal event times were not censored by other non-fatal event times. This trial was conducted in the United States and is of interest because of the observed marginally significant treatment effect (p=0.10), with an estimated hazard ratio of 0.90 (0.78, 1.02) for the effect of treatment on mortality. Although the results for the primary endpoint were not found to be significantly in favor treatment, secondary endpoints generally supported a beneficial treatment effect. Estimated marginal hazard ratios for all-cause hospitalization, MI, and heart transplant were 0.92, 0.52, and 0.69, respectively.

These marginal results for each of these endpoints are summarized in Table 3.1.

Table 3.1: Estimated marginal treatment effects with respect to component outcomes

Outcome	Events: Placebo Arm	Events: Bucindolol Arm	HR	p-value
Hospitalization	875	829	0.92 (0.84, 1.01)	0.08
MI	85	46	0.52 (0.36, 0.75)	<0.001
Heart Transplant	41	29	0.69 (0.43, 1.10)	0.12
Death	449	411	0.90 (0.78, 1.02)	0.10

It is reasonable to assume that these outcomes can be ordered by severity. Naturally, death may be considered the worst patient outcome, and any patient no longer living at time t should be assigned to the worst outcome category, in this case $E(t) = 4$. Among living patients, a heart transplant represents major surgery and may be considered an indication that a patient is nearing death. Thus, any patient alive at time t who has undergone a heart transplant would be classified as $E(t) = 3$. Among the remaining the patients, MI represents a relatively serious concern, and any patient who has experienced at least one MI prior to time t would be classified as $E(t) = 2$. Among those patients free of major cardiovascular complications (i.e., alive with no MI or transplant), we may further distinguish between those who have been admitted to the hospital for any reason ($E(t) = 1$), and those who remain completely event-free ($E(t) = 0$). The maximum follow-up time in the study was approximately 50 months with mean follow-up time of 24 months (Beta-Blocker Evaluation of Survival Trial Investigators, 2001). The censoring rate at t=1, 2, and 3 years was approximately 10%, 30%, and 65%, respectively. The estimated state probabilities, via the standard Kaplan-Meier curves, are plotted below in Figure 3.1, and the estimated state probabilities at followup times of 1, 2, and 3 years are shown in Table 3.2 below. We note that $P(E \leq k|Z = 1) > P(E \leq k|Z = 0)$ for all values of k at all three time points in the table, suggesting that patients receiving bucindolol are generally more likely to be in a healthier state than their counterparts at a given time. Furthermore, note that the set of comparisons between $P(E \leq k|Z = 1)$ and $P(E \leq k|Z = 0)$, $k = 0, \ldots K - 1$ represents a set of nested composite outcomes, evaluating the impact of treatment with respect to a) any clinical event (k=0), b) any major CV complication or death (k=1), c) heart transplant or death (k=2), and d) death (k=3).

Table 3.2: Estimated Patient Outcome Probabilities at 1, 2, and 3 years after treatment initiation

	Placebo	Bucindolol		Placebo	Bucindolol
$t = 1$ years					
$P(E \leq 0)$	0.487	0.519	$P(E = 0)$	0.487	0.519
$P(E \leq 1)$	0.831	0.856	$P(E = 1)$	0.345	0.336
$P(E \leq 2)$	0.849	0.862	$P(E = 2)$	0.018	0.006
$P(E \leq 3)$	0.858	0.868	$P(E = 3)$	0.019	0.006
$P(E \leq 4)$	1.000	1.000	$P(E = 4)$	0.142	0.132
$t = 2$ years					
$P(E \leq 0)$	0.312	0.351	$P(E = 0)$	0.312	0.351
$P(E \leq 1)$	0.681	0.720	$P(E = 1)$	0.369	0.369
$P(E \leq 2)$	0.701	0.730	$P(E = 2)$	0.021	0.010
$P(E \leq 3)$	0.720	0.745	$P(E = 3)$	0.018	0.014
$P(E \leq 4)$	1.000	1.000	$P(E = 4)$	0.280	0.255
$t = 3$ years					
$P(E \leq 0)$	0.195	0.237	$P(E = 0)$	0.195	0.237
$P(E \leq 1)$	0.546	0.606	$P(E = 1)$	0.351	0.369
$P(E \leq 2)$	0.575	0.619	$P(E = 2)$	0.029	0.014
$P(E \leq 3)$	0.604	0.635	$P(E = 3)$	0.029	0.016
$P(E \leq 4)$	1.000	1.000	$P(E = 4)$	0.396	0.365

Using the integrated general risk difference approach, we find that, after four years of followup, treated patients experienced a significant improvement in overall health, spending an expected 59.3 days spent in improved health relative to their untreated counterparts, with a 95% confidence interval of (10.9, 107.8) days, and corresponding p-value of 0.016. It should be noted that, by this measure, no significant benefit of treatment is detectable until nearly 1000 days (approximately 2.75 years) after initiation, and that furthermore, treatment was associated with significant harm through the first 4 months of follow-up with no positive estimated effect emerging until approximately 1 year after randomization. The estimated values and corresponding 95% confidence intervals for $IGRD(\cdot)$ are shown in Figure 3.2. For comparison, the difference in 4-year restricted mean survival times (with standard errors) corresponding to the individual Kaplan-Meier curves are 39.9 days (21.3), 55.7 days (21.1), 41.9 days (21.1), and 31.8 days (21.2), corresponding to $S_1(\cdot), S_2(\cdot), S_3(\cdot)$, and $S_4(\cdot)$, respectively.

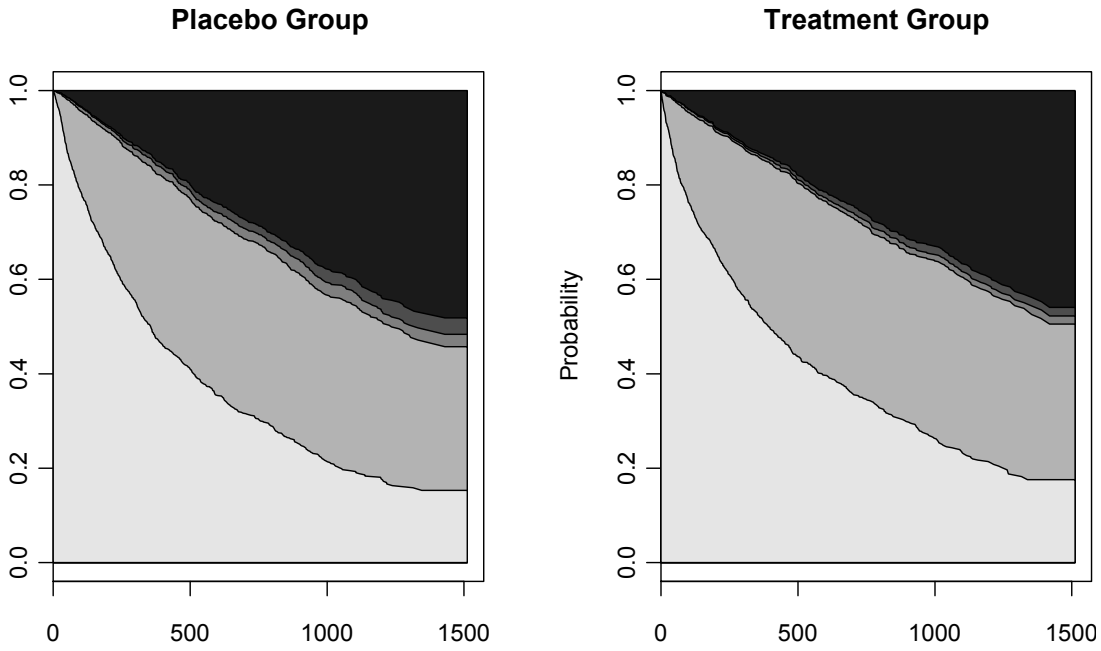


Figure 3.1: Estimated distribution of patient outcome status over the duration of BEST trial. Darker colors represent worsening health status (e.g. k=0 denoted by light gray; k=4 denoted by black.)

A similar trend is detected via the repeated ordinal regression approach. The model was fit at followup times of 3, 6, 12, 18, 24, 30, 36, and 48 months. Using the logit link, $g(p) = log(\frac{p}{1-p})$, a positive value of $\beta(t)$ indicates an increased probability of a patient being in a lower (healthier) state as a result of treatment. We again find some evidence of early harm in the treatment group, with $\hat{\beta}(t_1) = -0.09$, implying an estimated odds ratio of 0.91 for relationship between treatment and improved health status, though this estimated odds ratio is not significantly different from the null value 1. Under the assumption of a constant treatment effect over time, the resulting estimate is $\hat{\beta}(\cdot) = 0.09$ with 95% CI (-0.033, 0.215), implying a non-significant treatment benefit with corresponding odds ratio of 1.10 (0.97, 1.24), p=0.15. The estimates and 95% confidence intervals for each $\beta(t)$ are shown in Figure 3.3. The horizontal dashed line and gray region represent the estimate and confidence interval for the (assumed) constant value $\beta(\cdot)$.

Using the global model with the link $g(p) = log(-log(p))$, corresponding to the

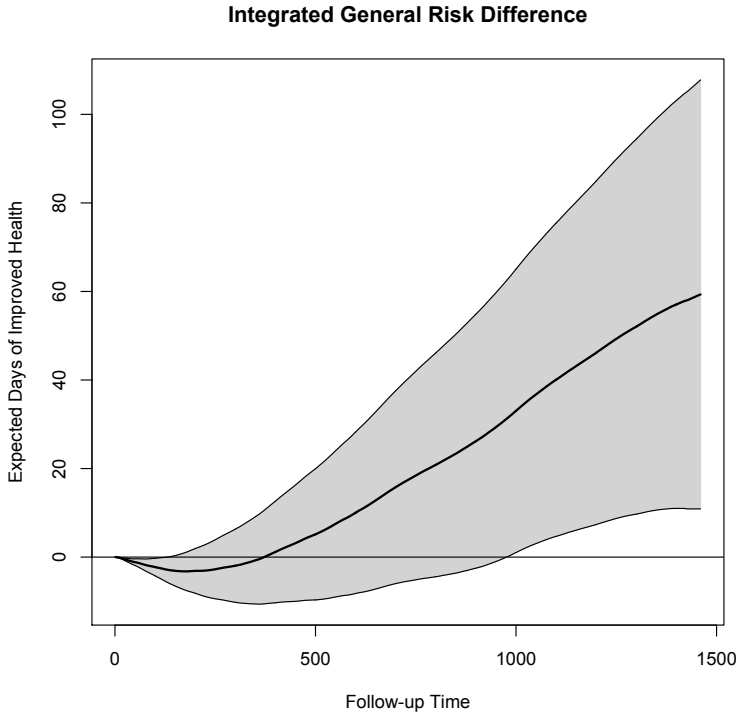


Figure 3.2: Estimated improvement in cumulative time spent in improved health, with pointwise 95% confidence intervals.

Wei-Lin-Weissfeld model, the estimated value $\hat{\beta}$ in this trial was -0.115, corresponding to a global hazard ratio of 0.89 associated with treatment. The robust variance procedure estimates a standard error of 0.050, with resulting 95% confidence intervals for $\beta = (-0.214, -0.017), HR = (0.81, 0.98)$, and associated p-value = 0.02, indicating a significant global benefit from treatment.

3.6 Discussion

In this paper, we discuss three potential approaches to the analysis of longitudinally assessed ordinal patient outcomes in clinical trials. Such ordinal outcomes are of particular interest for several reasons. The longitudinal nature of the data more easily allows for the assessment of changing treatment effects over time. The ordinal nature of the data allows for a

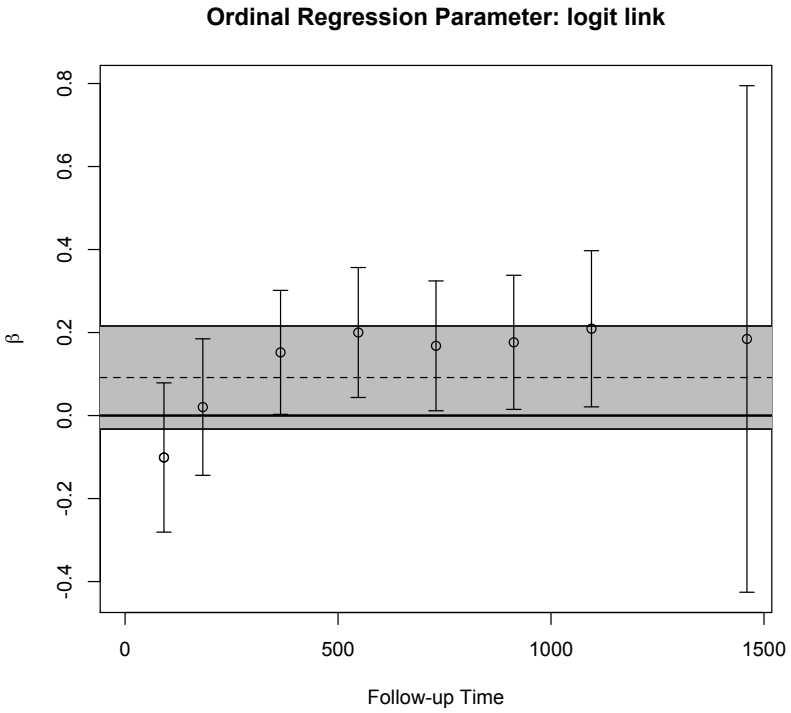


Figure 3.3: Estimated improvement in cumulative time spent in improved health, with pointwise 95% confidence intervals.

direct and efficient approach to incorporate many composite clinical outcomes into a single measure, and will provide more power to detect a significant clinical effect if the treatment improves patient health with respect to several related outcomes. The simultaneous assessment of multiple clinical outcomes often requires the imposition of utility weights to reflect the relative importance of each outcome with respect to the others. In many cases, these weights may be subjective and may differ between individual patients and/or clinicians. It is presumed that a ranking of such outcomes may be more readily agreed upon than a set of specific weights. In our example data set, we generally find that treatment with bucindolol is associated with significant improvement in long-term overall patient health, despite possible early negative effects during the first 6-12 months of treatment. These message seems to be consistent, regardless of the methods used.

Each of the three approaches described above, the integrated general risk difference

(IGRD), repeated ordinal regression, and the global survival model, has its own strengths and weaknesses. The IGRD approach, for example, requires no assumptions about the distribution of outcomes at a given follow-up time, and as such, does not require any particular link function or any other parametric assumptions. Furthermore, this approach easily allows for the treatment effect to change with time, and also provides an interpretable global summary measure in the form of expected time spent in improved health due to treatment. However, this approach does not easily allow for the introduction of covariates, beyond the binary treatment assignment, into the assessment of the treatment impact. Both the repeated ordinal regression and global survival models allow for such covariate effects to be built directly into their respective models. Unless further restrictions are placed on the repeated ordinal regression models, however, the estimated covariate effects may not be constant across analysis times. If the introduction of covariates into the model is solely for the purposes of improving the precision of the estimated treatment, then this will pose no problem. If one is instead interested in the overall effect of a particular set of covariates on patient outcomes across time, then the global survival model may be preferred. Additionally, it should be noted that standard software packages may be used to implement the repeated ordinal regression models with many commonly used link functions, while the global survival model is difficult to fit in standard software packages except in the special case mentioned above with $g(\cdot) = log(-log(\cdot))$ and a single terminal state. Additionally the global survival model requires the assumption that the treatment effect is constant over time, and the resulting parameter estimate may be interpreted as in a traditional Cox model, with $exp\{\hat{\beta}\}$ representing the overall hazard ratio for the instantaneous risk of a patient experiencing any decline in health.

Finally, it may not always be reasonable to impose a progressive disease model as we have assumed thus far. For example, patient health may improve over time or the criteria used to assign patients to outcome states may change over time. The correspondence between longitudinal cumulative link models and the survival function employed in nonparametric and semiparametric survival methods requires this progressive assumption, and so the global survival model will likely be impractical in such an unstructured setting. However the IGRD

and repeated ordinal regression approach can easily be adapted to such a setting. Further research is needed to develop methods that do not rely on a such progressive disease model.

3.7 Appendix

Alternate Derivation of $IGRD(t)$

It can be shown that the additional expected amount of time that patients in the treated group spend in a healthier state than patients in the untreated group is equivalent to the value of $IGRD(t)$. Let $X(t)$ and $Y(t)$ be the health statuses of a treated and an untreated patient, respectively. In addition, for specific patient histories $x(\cdot)$ and $y(\cdot)$, let $T(x, y)$ denote the duration of time, up to some follow-up time τ, spent with $x(t) < y(t)$. We are interested in $\mathbb{E}\{T(X,Y) - T(Y,X)\}$.

$$
\begin{aligned}
&\mathbb{E}\{T(X,Y) - T(Y,X)\} \\
=&\mathbb{E}\int_0^\tau \Big[I\{X(s) < Y(s)\} - I\{Y(s) < X(s)\}\Big] ds \\
=&\int_0^\tau \mathbb{E}\Big(\mathbb{E}\Big[I\{X(s) < y(s)\} \mid Y(s) = y(s)\Big]\Big) ds - \int_0^\tau \mathbb{E}\Big(\mathbb{E}\Big[I\{Y(s) < x(s)\} \mid X(s) = x(s)\Big]\Big) ds \\
=&\int_0^\tau \mathbb{E}\Big[P\Big\{X(s) < y(s) \mid Y(s) = y(s)\Big\}\Big] ds - \int_0^\tau \mathbb{E}\Big[P\Big\{Y(s) < x(s) \mid X(s) = x(s)\Big\}\Big] ds \\
=&\int_0^\tau \mathbb{E}\Big[\sum_k I\{Y(s) = k\} P\{X(s) < k\}\Big] ds - \int_0^\tau \mathbb{E}\Big[\sum_k I\{X(s) = k\} P\{Y(s) < k\}\Big] ds \\
=&\int_0^\tau \sum_k P\{Y(s) = k\} S_{1k}(s) ds - \int_0^\tau \sum_k P\{X(s) = k\} S_{0k}(s) ds \qquad (3.17) \\
=&\int_0^\tau \sum_k \{S_{0,k+1}(s) - S_{0k}(s)\} S_{1k}(s) ds - \int_0^\tau \sum_k \{S_{1,k+1}(s) - S_{1k}(s)\} S_{0k}(s) ds \\
=&\int_0^\tau \left[\sum_{k=1}^K \pi_{0,k}(s) \{\sum_{j=0}^{k-1} \pi_{1,j}(s)\} - \sum_{k=1}^K \pi_{1,k}(s) \{\sum_{j=0}^{k-1} \pi_{0,j}(s)\}\right] ds \\
=&IGRD(\tau)
\end{aligned}
$$

An inference procedure for order parameters utilizing confidence distribution random variables

Brian Claggett Min-ge Xie, and Lu Tian

Department of Biostatistics
Harvard School of Public Health

4.1 Introduction

Meta-analysis is a valuable tool for combining information from independent studies, especially when no single study is able to address the question at hand. The use of meta-analysis methods has grown substantially in recent years, with over 2000 papers per year published in PubMed, as of 2006 (Sutton and Higgins, 2008). Among these approaches, fixed-effects and random-effects models (particularly the DerSimonian-Laird approach) are two of the most commonly used models in meta-analysis. In practice, however, it is difficult, if not impossible, to verify the fundamental assumptions of these two models (i.e., one assumes homogeneous treatment effects across studies in a fixed-effects model or that the underlying study parameters are samples from a single (often normal) distribution in a random-effects model). Our question is: Can we develop a meta-analysis framework without taking a leap of faith to simply adopt these conventional assumptions? In this article, we develop such a framework to address this question and also provide a theoretical support and valid inference for related statistical problems.

Suppose that there are K independent studies whose underlying unknown parameters are denoted by $\theta_1, \ldots, \theta_K$. We make the minimal assumption that we only know that they are fixed unknown values, and any or all of them may or may not be equal to one another. In this context, similar to some nonparametric development, we may prefer to use, say, the median or other quantiles of the θ_i's as an overall quantification for the K studies. Furthermore, it may often be the case that either the minimum or maximum of a set of parameters is of particular interest. Finally, we may wish to construct an empirical function and make inference for the entire range of potential θ_i values. These considerations are all associated with the quantiles or the 'order statistics' of $\{\theta_1, \ldots, \theta_K\}$. Thus, we define the parameter of interest under this framework as the qth quantile (or the mth smallest) of the θ_i's:

$$\theta^{(m)} = \text{ the } m\text{th smallest } \theta_i,\ i = 1, \ldots, K. \tag{4.1}$$

We would like to make inference for $\theta^{(m)}$. Despite that we have defined the underlying

study parameters, we refer to this setup as a nonparametric framework in the sense that the assumptions made about models and/or distributions concerning the study parameters are minimal.

Without loss of generality, as argued in Xie et al. (2009) and Hall and Miller (2010), we assume that, from the ith study, we have a $\sqrt{n_i}$-consistent estimator of θ_i, say $\hat{\theta}_i$, where n_i is the sample size of the ith study. In addition, for simplicity and also following Xie et al. (2009) and Hall and Miller (2010), we assume further that the sample sizes $n_i \to \infty$ at the same rate and $\hat{\theta}_i \sim N(\theta_i, s_i^2)$ with the variance $s_i^2 = O(1/n_i)$, although the normality assumption can be relaxed. Let $\Theta = \{\theta_1, \ldots, \theta_K\}$ and $\hat{\Theta} = \{\hat{\theta}_1, \ldots, \hat{\theta}_K\}$. Also, let $N = \sum_{i=1}^K n_i$ and, without loss of generality, write $n_i = \lambda_i N$ for some positive constant λ_i's bounded away from zero, where $\sum_{i=1}^K \lambda_i = 1$. Our problem now, given the observed data $\{(\hat{\theta}_i, s_i, n_i),\ i = 1, 2, \ldots, K\}$, is how to make inference and construct confidence intervals for $\theta^{(m)}$ for any particular m.

This seemingly simple and innocent inference setup turns out to be associated with a well-known difficult problem. Hall and Miller (2010), who studied "the problem of constructing confidence intervals or hypothesis tests for extrema of parameters, for example of $\max\{\theta_1, \ldots, \theta_K\}$," stated that this type of problem is one of the "important problems where standard bootstrap estimators are not consistent, and where alternative approaches ... also face significant challenges." The difficult part is the unknown 'ties' and 'near ties' cases. Here, a near tie case is defined as

$$|\theta_j - \theta^{(m)}| = O(N^{-1/2}), \tag{4.2}$$

which is interpreted as that, based on current sample size $n_j = \lambda_j N$, a "near tie" parameter θ_j can not be distinguished from the parameter $\theta^{(m)}$; See Xie et al. (2009) and Hall and Miller (2010). In these cases, Hall and Miller (2010) stated that "the limiting distribution of $\max_{1 \le j \le K} \hat{\theta}_j$ might not be estimable by any method, be it the bootstrap or another approach." The approach recommended by Hall and Miller (2010) for this problem, as well as a set of more general forms of extreme parameters, was to construct a conservative confidence interval (test) by introducing a constant c_α to enlarge the usual confidence interval and use

bootstrap to estimate (tune) the constant c_α. Although the approach may be practical, it is conservative and fails to directly address the difficult problem of making inference on the extrema and other quantiles of the parameters.

In this paper, based on recent developments on confidence distributions (cf., a review article by Xie and Singh (2012)), we propose a new resampling method to deal with the inference problem for the extrema of the parameters and also, more generally, for any order statistics of the parameters. This new resampling method can be viewed as an extension of the well-studied and widely-used bootstrap method, but it enjoys a more flexible interpretation and manipulation. In the proposed method, we avoid the difficult problem of estimating the limiting distribution of $\hat{\theta}^{(m)}$. Rather, we directly construct an asymptotic confidence distribution for $\theta^{(m)}$, which can lead to asymptotically proper inference for any ordered parameters $\theta^{(m)}$. The rest of the paper is arranged as follows. In section 2, we introduce and review the idea of confidence distributions as frequentist distributional estimators, along with connections to the related bootstrap estimators. In section 3, we propose a general method for deriving an asymptotic confidence distribution for a particular $\theta^{(m)}$, which depends on a particular set of weights, and propose three reasonable weighting schemes, including the standard bootstrap estimator. In section 4, we discuss the properties of a set of weights which will guarantee appropriate asymptotic coverage, show how to construct weights that possess these properties, and discuss tuning approaches for finite-sample inference. In section 5, we present simulation results showing that our proposed weights provide appropriate coverage in diverse settings. In section 6, we illustrate our method using data from a recently published meta-analysis investigating the effect of an antioxidant on nephropathy.

4.2 CD-based Inference

4.2.1 Introduction to Confidence Distribution

In frequentist inference, we often use a single point of sample statistic (point estimator) or an sample-dependent interval (confidence interval) to estimate a parameter of interest. A confidence distribution (CD) is quite similar, but uses a sample-dependent distribution function, instead of a single point (point estimator) or an interval (confidence interval), to estimate the parameter of interest; cf., Xie and Singh (2012) and references therein. A confidence distribution has also been loosely referred to as a sample-dependent distribution function that can represent confidence intervals of all levels for the parameter of interest (see, e.g., Cox (1958); Efron (1993)). The concept has a long history, especially with its early interpretation associated with fiducial reasoning (see, e.g., Fisher (1973); Cox (2006)). In recent years, the confidence distribution concept has attracted a surge of renewed attention, and the recent developments have been based on a redefinition of the confidence distribution as a purely frequentist concept, without any fiducial reasoning. One nice aspect of this redefinition is that the confidence distribution is now a clean and coherent frequentist concept (similar to a point estimator) and is freed from those restrictive, if not controversial, constraints set forth by Fisher on fiducial distributions. In an invited review article, Xie and Singh (2012) provided a comprehensive review of the confidence distribution, including its history and a modern definition as well as its emerging new developments and its intertangled relationship with fiducial and Bayesian inference. A key aspect of the new developments is that a confidence distribution is "viewed as an estimator *for* the parameter of interest, instead of an inherent distribution *of* the parameter," which is different from the interpretation of a fiducial distribution or a Bayesian posterior. These developments on confidence distribution can be viewed as a part of recent developments of distributional inference in statistics, which include the concepts of generalized fiducial inference, belief function and objective Bayes. Its role in frequentist inference is similar to that of a posterior distribution in Bayesian inference.

A confidence distribution is a function of both the parameter and the random sample. It must satisfy certain requirements in order to provide appropriate inference for the parameter of interest. The following definition is formulated in Schweder and Hjort (2002); Singh et al. (2005). In the definition, Ξ is the parameter space of the unknown parameter of interest θ and $\mathcal{X}$ is the sample space corresponding to data $\mathbf{X}_n = \{X_1, \ldots, X_n\}$. Singh et al. (2005) demonstrated that this version of the confidence distribution definition is consistent with the classical definition which is compiled from confidence intervals of varying confidence levels (cf., Efron (1993)), and it is easier to use in many situations.

DEFINITION 2.1. *A function $H(\cdot) = H(\mathbf{X}_n, \cdot)$ on $\mathcal{X} \times \Xi \to [0, 1]$ is called a confidence distribution (CD) for a parameter θ, if it follows two requirements: R1) For each given $\mathbf{X}_n \in \mathcal{X}$, $H_n(\cdot)$ is a continuous cumulative distribution function on Θ; R2) At the true parameter value $\theta = \theta_0$, $H(\theta_0) \equiv H(\mathbf{X}_n, \theta_0)$, as a function of the sample $\mathbf{X}_n$, follows the uniform distribution $Unif[0, 1]$. Also, the function $H(\cdot)$ is an asymptotic CD (aCD), if the $Unif[0, 1]$ requirement is true only asymptotically and the continuity requirement on $H(\cdot)$ is dropped.*

Based on the definition, any sample-dependent distribution function on the parameter space can potentially be used to estimate the parameter, but the $Unif[0, 1]$ requirement in *R2* is imposed to ensure that the statistical inferences (e.g., point estimates, confidence intervals, p-values, etc.) derived from the confidence distribution have the desired frequentist properties. This practice of two requirements has an analog in point estimation: any single point (a real value or a statistic) on the parameter space can potentially be used to estimate a parameter, but we often impose restrictions so that the point estimator will have certain desired properties, such as unbiasedness, consistency, etc. As defined, the concept of confidence distribution is quite broad. It subsumes and unifies a wide range of examples, from regular parametric (fiducial distribution) examples to bootstrap distributions, significance (p-value) functions, normalized likelihood functions, and, in some cases, Bayesian priors and posteriors.

A simple example of a confidence distribution that has been broadly used in statistical practice is a bootstrap distribution. Efron (1998) explicitly stated that a bootstrap distribution is typically a "distribution estimator" and a "confidence distribution" function of the parameter that it targets. Singh et al. (2005, 2007) showed that a bootstrap distribution typically satisfies the definition of a confidence distribution or an asymptotic confidence distribution. In any situation where one can construct a bootstrap distribution, one can construct a confidence distribution or an asymptotic confidence distribution as well.

Another simple example, which is also used by Fisher (1930, 1973) to illustrate his fiducial function, is from the normal mean inference problem with sample $X_i \sim N(\mu, \sigma^2)$, $i = 1, \ldots, n$. The basic confidence distributions for μ are $\Phi(\sqrt{n}(\mu - \bar{X})/\sigma)$ when σ is known and $T_{n-1}(\sqrt{n}(\mu - \bar{x})/s)$ when σ is not known, and furthermore $\Phi(\sqrt{n}(\mu - \bar{X})/s)$ is an asymptotic confidence distribution when $n \to \infty$. Here, $\bar{X}$ and s^2 are the sample mean and variance, respectively, and T_{n-1} stands for the cumulative distribution function of the t-distribution with $n-1$ degrees of freedom. In other words, $N(\bar{X}, \sigma^2)$ is a "distribution estimator" of μ, when σ^2 is known. The distribution functions $T_{n-1}(\sqrt{n}(\mu - \bar{x})/s)$ or $N(\bar{X}, s^2)$ can be used to estimate μ, when σ^2 is not known. Similarly, in the context under consideration in this article, we can verify from Definition 2.1 that

$$H_i(t) = \Phi\left(\frac{t - \hat{\theta}_i}{s_i}\right) \tag{4.3}$$

satisfies the requirements of being an asymptotic confidence distribution, thus we can use a distribution estimator $N(\hat{\theta}_i, s_i^2)$ to estimate θ_i, for $i = 1, 2, \ldots, K$.

Since, for each given sample $\mathbf{X}_n$, a confidence distribution $H_n(\cdot)$ is a cumulative distribution function on the parameter space, we can construct a random variable ξ defined on $\mathcal{X} \times \Xi$ such that, conditional on the sample data, ξ has the distribution H. For example, let U be a $Unif[0,1]$ random variable independent of $\mathbf{X}_n$, then $\xi = H_n^{-1}(U) \sim H(\cdot)$, given $\mathbf{X}_n$. We call this random variable ξ a *CD random variable* (see, e.g., Singh et al. (2007); Xie and Singh (2012)).

DEFINITION 2.2. *We call $\xi = \xi_H$ a CD random variable associated with a confidence distribution H, if the conditional distribution of ξ given the data $\mathbf{X}_n$ is H.*

A CD random variable has a close association with a bootstrap estimator. In our example (4.3), a CD random variable ξ_i follows $\xi_i|\bar{x} \sim N(\hat{\theta}_i, s_i^2)$ and we have, asymptotically,

$$\frac{\xi_i - \hat{\theta}_i}{s_i}\,\Big|\,\hat{\theta}_i \;\sim\; \frac{\hat{\theta}_i - \theta_i}{s_i}\,\Big|\,\theta \qquad (\text{both} \sim N(0,1)). \tag{4.4}$$

This statement is exactly the same as the key justification for bootstrap, with ξ_i in place of the bootstrap sample mean $\hat{\theta}_i^*$. Thus, a CD random variable ξ can essentially be viewed as a model-based bootstrap estimator of θ_i. Indeed, Xie and Singh (2012) demonstrated under a very general setting that a CD random variable ξ is in essence the same as a bootstrap estimator or a simple linear transformation of a bootstrap estimator. This close connection between the CD random variable and a bootstrap estimator may inspire a possible view of treating the concept of confidence distribution as an extension of a bootstrap distribution, albeit the concept of confidence distribution is much broader. The connection and the well-developed theory of bootstrap distributions can help us to understand inference procedures involving confidence distributions and develop new methodologies. In this article, we utilize the CD random variable and develop a new simulation mechanism to broaden the applications of the standard bootstrap procedures. Since a CD random variable is not limited solely to use as a bootstrap estimator, this freedom allows us to utilize ξ more liberally, which in turn allows us to develop more flexible statistical approaches and inference procedures.

4.3 Proposed Methodology

As illustrated in (4.3), from the ith study we have a confidence distribution (CD) function $H_i(t) = \Phi((t - \hat{\theta}_i)/s_i)$ that can be used to estimate θ_i, for $i = 1, \ldots, K$. Denote by ξ_i the CD random variable corresponding to $H_i(t) = \Phi((t - \hat{\theta}_i)/s_i)$, i.e.,

$$\xi_i|\hat{\theta}_i, s_i^2 \sim N(\hat{\theta}_i, s_i^2), \quad \text{for } i = 1, \ldots, K. \tag{4.5}$$

Given a particular realized set of $\{\xi_i, i = 1, \ldots, K\}$ from each of the K studies, we consider the construction of a weighted average of ξ_i's:

$$\xi^* = \sum_{i=1}^{K} w_{i,(m)}\xi_i \Big/ \sum_{i=1}^{K} w_{i,(m)}. \tag{4.6}$$

We propose to make inference on $\theta^{(m)}$ based on ξ^*. In particular, we can easily simulate $\{\xi_i, i = 1, \ldots, K\}$ according to (4.5) and compute ξ^* according to (4.6). If we repeat this a large number of times, we can obtain a set of ξ^*'s, which represents a CD for the parameter $\theta^{(m)}$. Then, we can report the mean/median/mode of the ξ^*'s as a point estimate of $\theta^{(m)}$, and the empirical $(\alpha/2)100\%$ and $(1-\alpha/2)100\%$ quantiles of the ξ^*'s as the level $(1-\alpha)100\%$ confidence interval for $\theta^{(m)}$.

The proposed procedure is very simple. Naturally, different choices of the weights $w_{i,(m)}$ lead to different procedures, and each procedure's resulting validity depends on the choice of its weights. In particular, we consider in this paper the following potential choices of weights:

Choice 1:

$$w^{[1]}_{i,(m)} = \mathbf{1}\{\hat{\theta}_i = \hat{\theta}^{(m)}\}, \tag{4.7}$$

where $\mathbf{1}\{\cdot\}$ is an indicator function and $\hat{\theta}^{(m)}$ is the mth smallest $\hat{\theta}_i$.

Choice 2:

$$w^{[2]}_{i,(m)} = \mathbf{1}\{\xi_i = \xi^{(m)}\}, \tag{4.8}$$

where $\xi^{(m)}$ is the mth smallest ξ_i.

Choice 3:

$$w^{[3]}_{i,(m)} = \mathcal{K}\big(\xi_i - \xi^{(m)}, b_L, b_R\big) \tag{4.9}$$

where $\mathcal{K}$ is a kernel function, and b_L, b_R represent the left-side and right-side kernel bandwidths. Without loss of generality, we henceforth assume a rectangular kernel, such that

$$\mathcal{K}\big(\xi_i - \xi^{(m)}, b_L, b_R\big) = \mathbf{1}\{-b_L \leq (\xi_i - \xi^{(m)}) \leq b_R\}. \tag{4.10}$$

Written this way, it is easy to see that $w^{[2]}_{i,(m)}$ represents a special case of $w^{[3]}_{i,(m)}$ in which $b_L = b_R \equiv 0$.

Weights $w^{[1]}_{i,(m)}$ and $w^{[2]}_{i,(m)}$ both represent intuitively appealing ways of estimating and making inference on $\theta^{(m)}$. The use of $w^{[1]}_{i,(m)}$ is equivalent to using the CD (and resulting confidence interval) associated with the m^{th} ordered $\hat{\theta}$. The use of $w^{[2]}_{i,(m)}$ corresponds to the use of the distribution of the mth ordered ξ_i, and is therefore equivalent to the conventional bootstrap estimator of $\theta^{(m)}$, as discussed in Hall and Miller (2010). Despite these intuitively attractive qualities, we will show that both sets of weights may lead to undesirable properties, depending on the true nature of the data, while our third option is flexible enough to appropriately handle a variety of scenarios while maintaining appropriate coverage levels, and in many cases, offering narrower confidence intervals than those obtained by the other weighting schemes. In the following section, we show that there is a very simple requirement for any given weighting scheme that allows for the use of ξ^* for asymptotically valid inference for $\theta^{(m)}$. Namely, $w_{i,(m)}$ must converge to a positive constant if θ_i belongs to the tie or near tie set of $\theta^{(m)}$, as defined below, and zero otherwise. We will show that this requirement is not satisfied by $w^{[1]}_{i,(m)}$ or $w^{[2]}_{i,(m)}$, but is satisfied by $w^{[3]}_{i,(m)}$ when $(b_L, b_R) = O(N^{-\delta}), \delta \in (0, \frac{1}{2})$.

4.4 Theoretical results

First, let us formally define the tie and near tie sets. The same definition has also been utilized in Xie et al. (2009); Hall and Miller (2010). In particular, we denote by

$$\Theta^{(m)}_{\mathcal{T}} = \{j : \theta_j = \theta^{(m)}, j = 1, \ldots, K\} \tag{4.11}$$

the "tie set" of $\theta^{(m)}$, representing the set of all θ's which are equal to the parameter of interest. We also denote by

$$\Theta^{(m)}_{\mathcal{N}} = \{j : |\theta_j - \theta^{(m)}| = O(N^{-1/2}), j = 1, \ldots, K\} \tag{4.12}$$

the "near tie" of $\theta^{(m)}$. The interpretation of the "near tie" definition is that, based on current sample size n_i, a "near tie" parameter θ_i cannot be distinguished from the target parameter $\theta^{(m)}$. An equivalent expression is that, for any $j \in \Theta_{\mathcal{N}}^{(m)}$, $(\hat{\theta}_j - \hat{\theta}^{(m)}) - (\theta_j - \theta^{(m)}) \neq o_p(|\theta_j - \theta^{(m)}|)$, which means that the difference between θ_j and $\theta^{(m)}$ is not of greater order than the standard error of its estimator. Throughout the paper, we assume that both $\Theta_{\mathcal{T}}^{(m)}$ and $\Theta_{\mathcal{N}}^{(m)}$ are completely unknown other than that they contain at least one member $\theta^{(m)}$. Thus, without loss of generality, we can assume the number of studies in the tie set $|\Theta_{\mathcal{T}}^{(m)}| \geq 1$. The "near tie" case is much broader than the tie case: $\Theta_{\mathcal{T}}^{(m)} \subseteq \Theta_{\mathcal{N}}^{(m)}$. So, we also have the number of studies in the near tie set $|\Theta_{\mathcal{N}}^{(m)}| \geq |\Theta_{\mathcal{T}}^{(m)}| \geq 1$.

We present next a set of theoretical results using the more general near tie setup. All results remain valid if $\Theta_{\mathcal{N}}^{(m)}$ is replaced by $\Theta_{\mathcal{T}}^{(m)}$.

4.4.1 Asymptotic theorem and properties of proposed weighing schemes

The following set of asymptotic results suggest that ξ^* may be used to make inference for $\theta^{(m)}$, if weights are chosen carefully. In the theorem, Ξ is the parameter space of $\theta^{(m)}$.

THEOREM 4.1. *Suppose that we can prove that a set of weights possesses the following property.*

$$\lim_{n\to\infty} w_{i,(m)} = \begin{cases} c_i & \text{if } i \in \Theta_{\mathcal{N}}^{(m)}, \\ 0 & \text{if } i \notin \Theta_{\mathcal{N}}^{(m)}, \end{cases} \quad \text{for } i = 1, 2, \ldots, K \tag{4.13}$$

for some constants $c_i > 0$. Then, as $N \to \infty$, we have the following:

(i)

$$\sum_{i=1}^{K} w_{i,(m)}\hat{\theta}_i \Big/ \sum_{i=1}^{K} w_{i,(m)} = \theta^{(m)} + o_p(1) \quad \text{and} \quad \sum_{i=1}^{K} w_{i,(m)}^2 s_i^2 \Big/ \Big\{ \sum_{i=1}^{K} w_{i,(m)} \Big\}^2 = \{s^{(m)}\}^2 + o_p(1), \tag{4.14}$$

where $\{s^{(m)}\}^2 = \sum_{i\in\Theta_{\mathcal{N}}^{(m)}} c_i^2 s_i^2 / \{ \sum_{i\in\Theta_{\mathcal{N}}^{(m)}} c_i \}^2$. Furthermore,

$$\frac{\xi^* - \sum_{i=1}^K w_{i,(m)}\hat{\theta}_i / \sum_{i=1}^K w_{i,(m)}}{\sqrt{\sum_{i=1}^K w_{i,(m)}^2 s_i^2 / \{\sum_{i=1}^K w_{i,(m)}\}^2}}\Bigg|\hat{\Theta} \sim \frac{\sum_{i=1}^K w_{i,(m)}\hat{\theta}_i / \sum_{i=1}^K w_{i,(m)} - \theta^{(m)}}{\sqrt{\sum_{i=1}^K w_{i,(m)}^2 s_i^2 / \{\sum_{i=1}^K w_{i,(m)}\}^2}}\Bigg|\Theta, \tag{4.15}$$

both converging asymptotically to a N(0, 1) distribution.

(ii) Denote by

$$H_*(t) = P(\xi^* \leq t|\hat{\Theta}), \quad \textit{for any } t \in \Xi. \tag{4.16}$$

When $t = \theta^{(m)}$*, we have* $H_*(\theta^{(m)}) \to Unif[0,1]$*, in distribution, and therefore* $H_*(\theta)$ *is an aCD for* $\theta^{(m)}$*.*

A proof of the theorem is provided in Appendix.

The function $H_*(t)$ is a cumulative distribution function on the parameter space Ξ and it also depends on the sample observations $\hat{\Theta}$. Definition 2.1 suggests that, when our weight choice satisfies the requirement (4.13), $H_*(\theta) = Pr(\xi^* \leq \theta|\hat{\Theta})$ is an aCD for $\theta^{(m)}$. Based on the development on CDs (see, e.g., Singh et al. (2007); Xie and Singh (2012)), it subsequently ensures asymptotically valid inference, including point estimation, confidence intervals, p-values, etc., regarding $\theta^{(m)}$. Thus, in this case, we can rely on ξ^* to provide valid inference for $\theta^{(m)}$ asymptotically.

The remaining question is whether any of the three sets of weight choices in Section 3 satisfy the requirement (4.13) and, if they do, under which conditions. Since the asymptotic properties of each of the proposed weighted estimators depend on the true unknown values of Θ, we start with the simplest setting of no ties and move on to the more complicated settings of ties and near ties, including the particularly difficult case in which the presence of such ties or near ties to $\theta^{(m)}$ cannot easily be determined.

The 'no tie' case is the case in which $|\Theta_{\mathcal{N}}^{(m)}| = |\Theta_{\mathcal{T}}^{(m)}| = 1$. In particular, this refers to the case that $\Theta_{\mathcal{N}}^{(m)}$ and $\Theta_{\mathcal{T}}^{(m)}$ have only one element, and the rest of θ_j, $j \notin \Theta_{\mathcal{N}}^{(m)}$, satisfy the

condition that

$$d_m N^{1/2} \to \infty, \quad \text{where} \quad d_m = \min_{j \notin \Theta_{\mathcal{N}}^{(m)}} \left| \theta_j - \theta^{(m)} \right| \tag{4.17}$$

is the minimal distance between the θ_j's in and outside the near tie set $\Theta_{\mathcal{N}}^{(m)}$. The condition (4.17) is in fact weaker than and covers the conventional assumption of no ties, in which $\theta_1, \theta_2, \ldots, \theta_K$ are unknown but distinct constant parameters. In this case, $d_m = \min_{\theta_j \neq \theta^{(m)}} |\theta_j - \theta^{(m)}| \geq c_o = \min_{i \neq j} |\theta_i - \theta_j|$ which is a positive constant bounded away from zero. There may or may not be ties among the remaining θ_j's, but this is irrelevant to the problem at hand in making inference for $\theta^{(m)}$.

Lemma 1 below states that, under the above no tie condition, all the three choices of weights listed in Section 3 satisfy the condition in (4.13). A proof is given in the Appendix.

LEMMA 4.1 (ANY WEIGHT; NO TIE CASE). *Suppose that* $|\Theta_{\mathcal{N}}^{(m)}| = |\Theta_{\mathcal{T}}^{(m)}| = 1$ *and also Condition (4.17) holds. For* $s = 1, 2$, *we have*

$$\lim_{N \to \infty} w_{i,(m)}^{[s]} = \begin{cases} 1 & \text{if } \theta_i = \theta^{(m)}, \\ 0 & \text{if } \theta_i \neq \theta^{(m)}, \end{cases} \quad \text{for } i = 1, 2, \ldots, K. \tag{4.18}$$

Furthermore, if we use $w_{i,(m)}^{[3]}$ *with* $b_L, b_R \propto \tau_N$, *where* $\tau_N / d_m \to 0$, *and* $\tau_N \sqrt{N} \to \infty$, *then (4.18) also holds for* $w_{i,(m)}^{[3]}$.

Accompanying Theorem 4.1, we can infer from the lemma that in the no tie case, we can implement the proposed approach using any of the three weight to make asymptotically valid inference for $\theta^{(m)}$. In fact, since (4.18) holds for all $s = 1, 2, 3$, it is easy to verify, following the proof of Theorem 4.1, that the inference based on these three different choices of weights are asymptotically equivalent.

The problem in the tie or the near tie case is more complicated. In this case, the weights $w_{i,(m)}^{[1]}$ or $w_{i,(m)}^{[2]}$ for $i \in \Theta_{\mathcal{T}}^{(m)}$ or $\Theta_{\mathcal{N}}^{(m)}$ converge to random quantities, rather than some constants c_i. We provide below a very simple example in a special case to illustrate the phenomenon.

EXAMPLE 4.1 (COUNTEREXAMPLE FOR $w_{i,(m)}^{[1]}$ OR $w_{i,(m)}^{[2]}$ IN A SIMPLE TIE CASE). Without loss of generality, consider a very simple example of a special case with $K = 2$ and $\theta_1 \equiv \theta_2$. For $m = 1$, $\Theta_{\mathcal{T}}^{(m)} = \Theta_{\mathcal{N}}^{(m)} = \{1, 2\}$ but $w_{1,(m)}^{[1]} = 1 - w_{2,(m)}^{[1]} = \mathbf{1}\{\hat{\theta}_1 = \min(\hat{\theta}_1, \hat{\theta}_2)\}$ is a binary random variable that equals 1 with probability $P\{\hat{\theta}_1 \leq \hat{\theta}_2\} = 1 - P\{\hat{\theta}_2 \leq \hat{\theta}_1\} = 0.5$. Thus, both $w_{1,(m)}^{[1]}$ and $w_{2,(m)}^{[1]}$ are (dependent) Bernoulli random variables, each with $p = 0.5$, therefore violating (4.13).

Similarly, for $m = 1$, $w_{1,(m)}^{[2]} = 1 - w_{2,(m)}^{[2]} = \mathbf{1}\{\xi_1 = \min(\xi_1, \xi_2)\}$ is a binary random variable that equals 1 with probability $P\{\xi_1 \leq \xi_2\} = E\big[P\{\xi_1 \leq \xi_2 | \hat{\Theta}\}\big] = E\big[\Phi(\{\hat{\theta}_2 - \hat{\theta}_1\}/\{s_1^2 + s_2^2\}^{1/2})\big] = 0.5$. Again, both $w_{1,(m)}^{[2]}$ and $w_{2,(m)}^{[2]}$ are (dependent) Bernoulli random variables, each with $p = 0.5$, also violating (4.13).

In the case of more than two ties with either $|\Theta_{\mathcal{T}}^{(m)}| > 2$ or $|\Theta_{\mathcal{N}}^{(m)}| > 2$, the weights $w_{i,(m)}^{[1]}$ or $w_{i,(m)}^{[2]}$ for $i \in \Theta_{\mathcal{T}}^{(m)}$ or $\Theta_{\mathcal{N}}^{(m)}$ still converge to random quantities, rather than constants. The patterns are similar to, but more complicated than, that discussed in the case of $|\Theta_{\mathcal{T}}^{(m)}| = 2$ in Example 4.1. Clearly, neither $w_{i,(m)}^{[1]}$ nor $w_{i,(m)}^{[2]}$ satisfies the requirement (4.13) in this case, thus we can no longer ensure that the results from Theorem 4.1 are valid. Our simulation results confirm that these two sets of weights perform poorly in situations with ties or near ties. Poor performance of the standard bootstrap procedure, which corresponds to the use of the second sets of weights $w_{i,(m)}^{[2]}$, was also reported by Hall and Miller (2010).

In contrast, if we use $w_{i,(m)}^{[3]}$ with $b_L, b_R \propto \tau_N$, where $\tau_N/d_m \to 0$ and $\tau_N\sqrt{N} \to \infty$, then we can show that (4.13) is satisfied, provided that Condition (4.17) holds. In fact, the requirement (4.13) is satisfied by $w_{i,(m)}^{[3]}$ in any case, regardless of whether or not any ties or near ties exist, and regardless of whether or not their existence can be determined from the data . We summarize the result in the following lemma, together with the result for a slightly modified $w_{i,(m)}^{[3]}$ choice:

$$\widetilde{w}_{i,(m)}^{[3]} = w_{i,(m)}^{[3]} \Big/ s_i. \tag{4.19}$$

A proof can be found in the Appendix.

LEMMA 4.2 (WEIGHT $w^{[3]}_{i,(m)}$; ANY CASE). *Suppose that Condition (4.17) holds and we use* $w^{[3]}_{i,(m)}$ *with* $b_L, b_R \propto \tau_N$, *where* $\tau_N/d_m \to 0$, *and* $\tau_N\sqrt{N} \to \infty$. *For any* $1 \le |\Theta^{(m)}_{\mathcal{T}}| \le |\Theta^{(m)}_{\mathcal{N}}| \le K$, *we have*

$$\lim_{N\to\infty} w^{[3]}_{i,(m)} = \begin{cases} 1 & \text{if } i \in \Theta^{(m)}_{\mathcal{N}}, \\ 0 & \text{if } i \notin \Theta^{(m)}_{\mathcal{N}}, \end{cases} \quad \text{and} \quad \lim_{N\to\infty} \widetilde{w}^{[3]}_{i,(m)} = \begin{cases} 1/s_i & \text{if } i \in \Theta^{(m)}_{\mathcal{N}}, \\ 0 & \text{if } i \notin \Theta^{(m)}_{\mathcal{N}}, \end{cases} \tag{4.20}$$

for $i = 1, 2, \ldots, K$.

This lemma, together with Theorem 4.1, provides a theoretical support to use the weighted sum of CD random variables ξ^* to make inference for $\theta^{(m)}$ in all cases, if either $w^{[3]}_{i,(m)}$ or $\widetilde{w}^{[3]}_{i,(m)}$ is used. From (4.20), only studies inside the tie and near tie set will be included for making inference and the studies outside the tie set are filtered out, asymptotically. Thus, making inference using the proposed method with $w^{[3]}_{i,(m)}$ is asymptotically equivalent to using the average of the $\hat{\theta}_i$ in the tie set (if we were to know the true tie set). When s_i's or $\lambda_i = n_i/N$'s are heteroscedastic, the modified version $\widetilde{w}^{[3]}_{i,(m)}$ could be used to improve the efficiency and power of the inference. In any case, as long as there is a separation between the studies not tied with $\theta^{(m)}$ and those tied with $\theta^{(m)}$ as quantified in Condition (4.17), our proposal provides a class of approaches that can lead us to asymptotically correct inference. Further details will be discussed in Section 4.2 on the tuning of the kernel widths. Note that, Condition (4.17) is much weaker than those assumptions imposed in the conventional fixed-effects and random effects models.

4.4.2 Tuning the bandwidth parameters

Let us first define the bandwidth parameters $b_L = \tau_N \cdot c_L, b_R = \tau_N \cdot c_R$, where $\tau_N = O(N^{-\delta})$ and $c_L, c_R = O(1)$. Throughout, we will use $\tau_N = (\sigma)(s^{(m)}/\sigma)^{1/2}$, where $s^{(m)}$ is the standard error associated with $\hat{\theta}^{(m)}$ and σ is reasonable maximum value for the tuning parameter, such as $\sigma = \sqrt{\frac{\sum_i s_i^2 * n_i}{K}}$. This particular formulation of σ ensures that $s^{(m)}/\sigma \approx n_{(m)}^{-1/2}$, and that $(s^{(m)}/\sigma)^{1/2} > (s^{(m)}/\sigma)$. Note that $\sigma = O(1), s^{(m)} = O(N^{-1/2})$, and so $\delta = 1/4$.

While we can guarantee that $w^{[3]}_{i,(m)}$ will provide appropriate asymptotic inference regardless of the selection of tuning parameters (c_L, c_R), it is important in practice to be able to select an appropriate value for the tuning parameters (c_L, c_R) to ensure good finite sample performance. We note that for sufficiently large values of $(c_L,\ c_R)$, our inference will mimic a fixed-effects analysis, which is only reasonable under the assumption that $|\Theta^{(m)}_T| = K$. On the other hand, when these bandwidth values are equal to 0, our weights are identical to $w^{[2]}_{i,(m)}$, which we have shown to be asymptotically valid only when $|\Theta^{(m)}_T| = 1$. Thus the tuning bandwidth values should be relatively large when ties are present and relatively small when no ties are present. Generally, we refer to $\hat{\theta}_j : \{\hat{\theta}_j < \hat{\theta}^{(m)}, \theta_j = \theta^{(m)}\}$ as "left side ties" and $\hat{\theta}_j : \{\hat{\theta}_j > \hat{\theta}^{(m)}, \theta_j = \theta^{(m)}\}$ as "right side ties". To this end, we attempt to detect the presence or absence of left side ties and/or right side ties by observing the behavior of the realized values of the CD random variables ξ.

In general, we will simulate some large number, R, of samples of our CD random variables $\{\xi\}$, and we may denote the r^{th} sampled value corresponding to $\hat{\theta}^{(i)}$ as $\xi_{i,r}$, the r^{th} collection of sampled values as $\{\xi_r\} = \{\xi_{i,r}, i = 1, \ldots, K\}$, and $\xi^{(m)}_r$ as the m^{th} smallest value of the sampled vector $\{\xi_r\}$.

We then define the following terms for the purposes of determining the presence or absence of tied θ values.

Let $\hat{\pi}_i = \frac{\sum_{r=1}^{R} \mathbf{1}\{\xi^{(\mathbf{m})}_{\mathbf{r}} = \xi_{\mathbf{i},\mathbf{r}}\}}{R}$, and $\hat{\mathcal{R}}_i = \mathbf{1}\{\hat{\pi}_i > 0.001\}$. Then $\hat{T} = \sum_i \hat{\mathcal{R}}_i$, $\hat{T}_L = \sum_{i<m} \hat{\mathcal{R}}_i$, $\hat{T}_R = \sum_{i>m} \hat{\mathcal{R}}_i$, $\hat{\pi}_L = \sum_{i<m} \hat{\pi}_i$, and $\hat{\pi}_R = \sum_{i>m} \hat{\pi}_i$. Noting that $\sum_i \hat{\pi} = 1$, and that $\hat{T} = (\hat{T}_L = \hat{T}_R + 1)$ gives an estimate of the maximum size of the tie set $|\Theta^{(m)}_T|$, both $\hat{\pi}_L$ and $\hat{T}_L/\hat{T}$ are values between 0 and 1 which provide information about the presence of ties on the left side of $\hat{\theta}^{(m)}$. The geometric mean of these two values $c^*_L = \sqrt{\hat{\pi}_L \cdot \hat{T}_L/\hat{T}}$ then provides a reasonable summary of the evidence regarding the existence and influence of left side ties, and similarly $c^*_R = \sqrt{\hat{\pi}_R \cdot \hat{T}_R/\hat{T}}$ on the right.

Empirically, in small sample settings where it is quite difficult to determine whether or

not there are ties, we find that (c_R^*, c_L^*) may under-smooth the data when $|\Theta_T^{(m)}|$ is quite large relative to K, and conversely, may over-smooth the data when $|\Theta_T^{(m)}| = 1$. In order to reflect the true uncertainty as to the size of the tie set, we propose to induce additional randomness into the smoothing procedure by utilizing $c_R = u \cdot c_R^*, c_L = u \cdot c_L^*$, where $u \sim Unif(\frac{1}{7\hat{T}K}, \frac{7\hat{T}}{K})$. Because $1 \leq \hat{T} \leq K$, we see that $\frac{1}{7K^2} \leq u \leq 7$, and therefore, $\frac{1}{7K^2} \leq c_L \leq 7$, $\frac{1}{7K^2} \leq c_R \leq 7$ and so u does not influence the convergence rate of the bandwidth parameters b_R, b_L, and only serves to induce additional variability into the resulting weighted averages ξ^*.

Thus the algorithm for obtaining the CD used to estimate $\theta^{(m)}$ is as follows:

1. Calculate τ_N using observed $(\hat{\theta}, \hat{s}^2, n)$ data.
2. Generate $\{\xi_{i,r}, r = 1, \ldots, R\}$ for each study from $N(\hat{\theta}^{(i)}, \hat{s}^{2(i)})$
3. For each vector $\{\xi_r\}$, determine which study i is associated with $\xi_r^{(m)}$.
4. Use these counts to calculate $\hat{\pi}_i$ for each study, and functions thereof $(\hat{\mathcal{R}}_i, \hat{T}, \hat{T}_L, \hat{T}_R, \hat{\pi}_L, \hat{\pi}_R, c_L^*, c_R^*)$.
5. For each r in $1, \ldots, R$
   (a) Generate u_r from $Unif(\frac{1}{7\hat{T}K}, \frac{7\hat{T}}{K})$
   (b) $\xi_r^* = \frac{\sum_i \xi_{i,r} \mathbf{1}\{-u_r\tau_N c_L^* \leq (\xi_{i,r} - \xi_r^{(m)}) \leq u_r \tau_N c_R^*\}}{\sum_i \mathbf{1}\{-u_r\tau_N c_L^* \leq (\xi_{i,r} - \xi_r^{(m)}) \leq u_r \tau_N c_R^*\}}$
6. The CD for $\theta^{(m)}$ is approximated by the empirical distribution $\hat{H}_{\xi^*}(\theta)$, and a $(1 - \alpha)100\%$ confidence interval can be estimated by $(\xi^*_{R(\alpha/2)}, \xi^*_{R(1-\alpha/2)})$.

4.5 Simulations

In order to demonstrate both small and large sample properties of our proposed estimator under different scenarios, we generate random data $X_{ij} \sim N(\theta_i, 1)$, with $\theta_i, i \in \{1, 2, \ldots, K\}, 1 \leq j \leq n_i$, taking different values according to the particular scenario:

1. Ties: $\theta_i = 0 \; \forall i$

2. Uniform: $\theta_i = \frac{i}{K+1}$

3. Normal: $\theta_i = \Phi^{-1}(\frac{i}{K+1})$

For each scenario, we consider $K = 7$ or $K = 21$, and we let the sample size from each study $n_i = 40$, 400, or 4000. Using 500 simulated data sets for each setting, we show the coverage and median width of the nominal 95% confidence interval.

We consider each of the three methods proposed in Section 2. The results are shown below. Because each set of $\{\Theta\}$ is symmetric, we need not show results for each ordered θ_i. In particular, the coverage and median interval width for any $\theta^{(k)}$ will be identical to that for $\theta^{(K+1-k)}$.

For our proposed method using kernel smoothing, the results shown use the tuning procedure described in the previous section with R=1000 random samples drawn from each study's confidence distribution. Simulation results are shown below.

We first note that Method 1 will always return confidence intervals of equal or greater width than those returned by Method 2. Correspondingly, we find many settings in which the coverage of Method 2 is far below the nominal level (e.g. the Ties setting, the Uniform setting with $n_i = 40$). In almost all of these settings (except for $m = 1, 2$ in the Ties case), Method 1 will provide appropriate, but conservative, confidence intervals. Our proposed Method 3, on the other hand, is shown to have appropriate coverage levels in all settings, as well as noticeably narrower confidence interval widths relative to Method 1 in nearly all cases. Relative to the bootstrap estimator (Method 2), the intervals from our proposed method are narrower, in the fixed-effects setting, for the few cases in which the bootstrap estimator provides appropriate coverage, and the interval widths are similar (and asymptotically equal) to those from Method 2 in the uniform and normal settings.

Table 4.1: Simulation results for settings with K=7

Setting Ties	$n_i = 40$ m	Method 1 Cov.	Method 1 Width	Method 2 Cov.	Method 2 Width	Method 3 Cov.	Method 3 Width	$n_i = 400$ m	Method 1 Cov.	Method 1 Width	Method 2 Cov.	Method 2 Width	Method 3 Cov.	Method 3 Width	$n_i = 4000$ m	Method 1 Cov.	Method 1 Width	Method 2 Cov.	Method 2 Width	Method 3 Cov.	Method 3 Width
Uniform																					
	1	**0.964**	0.612	**0.912**	0.501	**0.958**	0.658	1	**0.954**	0.194	**0.958**	0.186	**0.944**	0.195	1	**0.952**	0.062	**0.952**	0.062	**0.952**	0.062
	2	**0.992**	0.610	**0.990**	0.421	**0.970**	0.541	2	**0.962**	0.196	**0.970**	0.175	**0.956**	0.193	2	**0.930**	0.062	**0.930**	0.062	**0.930**	0.062
	3	**0.994**	0.612	**0.996**	0.396	**0.976**	0.463	3	**0.964**	0.195	**0.974**	0.174	**0.942**	0.197	3	**0.936**	0.062	**0.936**	0.062	**0.936**	0.062
	4	**0.998**	0.614	**0.998**	0.388	**0.982**	0.422	4	**0.960**	0.194	**0.970**	0.174	**0.948**	0.197	4	**0.934**	0.062	**0.934**	0.062	**0.934**	0.062
Normal																					
	1	**0.942**	0.609	**0.944**	0.588	**0.942**	0.599	1	**0.952**	0.194	**0.952**	0.194	**0.952**	0.194	1	**0.952**	0.062	**0.952**	0.062	**0.952**	0.062
	2	**0.938**	0.616	**0.950**	0.544	**0.942**	0.584	2	**0.954**	0.196	**0.954**	0.196	**0.954**	0.196	2	**0.930**	0.062	**0.930**	0.062	**0.930**	0.062
	3	**0.956**	0.607	**0.964**	0.520	**0.950**	0.576	3	**0.942**	0.195	**0.942**	0.195	**0.942**	0.195	3	**0.936**	0.062	**0.936**	0.062	**0.936**	0.062
	4	**0.966**	0.617	**0.970**	0.520	**0.954**	0.578	4	**0.942**	0.194	**0.942**	0.194	**0.942**	0.194	4	**0.934**	0.062	**0.934**	0.062	**0.934**	0.062

Table 4.2: Simulation results with K=21

	$n_i = 40$	Method 1		Method 2		Method 3		$n_i = 400$	Method 1		Method 2		Method 3		$n_i = 4000$	Method 1		Method 2		Method 3	
	m	**Cov.**	Width	**Cov.**	Width	**Cov.**	Width	m	**Cov.**	Width	**Cov.**	Width	**Cov.**	Width	m	**Cov.**	Width	**Cov.**	Width	**Cov.**	Width
Ties																					
	1	**0.580**	0.614	**0.000**	0.400	**0.934**	0.566	1	**0.584**	0.196	**0.000**	0.125	**0.958**	0.161	1	**0.564**	0.061	**0.000**	0.040	**0.974**	0.045
	2	**0.864**	0.605	**0.000**	0.298	**0.980**	0.426	2	**0.896**	0.195	**0.000**	0.094	**0.978**	0.122	2	**0.902**	0.062	**0.000**	0.029	**0.978**	0.033
	3	**0.974**	0.618	**0.000**	0.256	**0.986**	0.358	3	**0.982**	0.195	**0.000**	0.081	**0.978**	0.102	3	**0.982**	0.062	**0.000**	0.026	**0.978**	0.027
	4	**0.990**	0.615	**0.004**	0.233	**0.986**	0.308	4	**0.994**	0.195	**0.002**	0.074	**0.978**	0.088	4	**0.994**	0.062	**0.000**	0.023	**0.982**	0.023
	5	**0.998**	0.608	**0.034**	0.217	**0.982**	0.274	5	**1.000**	0.194	**0.030**	0.069	**0.978**	0.077	5	**0.998**	0.062	**0.028**	0.022	**0.980**	0.020
	6	**1.000**	0.610	**0.144**	0.207	**0.986**	0.243	6	**1.000**	0.195	**0.158**	0.066	**0.980**	0.068	6	**1.000**	0.062	**0.144**	0.021	**0.978**	0.018
	7	**1.000**	0.612	**0.388**	0.200	**0.986**	0.218	7	**1.000**	0.195	**0.398**	0.064	**0.980**	0.061	7	**1.000**	0.062	**0.374**	0.020	**0.978**	0.017
	8	**1.000**	0.609	**0.670**	0.195	**0.986**	0.199	8	**1.000**	0.194	**0.678**	0.062	**0.976**	0.056	8	**1.000**	0.062	**0.706**	0.020	**0.974**	0.016
	9	**1.000**	0.619	**0.890**	0.192	**0.986**	0.184	9	**1.000**	0.195	**0.904**	0.061	**0.978**	0.052	9	**1.000**	0.061	**0.906**	0.019	**0.972**	0.015
	10	**1.000**	0.609	**0.978**	0.190	**0.984**	0.175	10	**1.000**	0.195	**0.976**	0.061	**0.972**	0.050	10	**1.000**	0.062	**0.978**	0.019	**0.970**	0.015
	11	**1.000**	0.610	**0.994**	0.189	**0.986**	0.172	11	**1.000**	0.195	**0.994**	0.060	**0.972**	0.050	11	**1.000**	0.062	**0.996**	0.019	**0.966**	0.015
Uniform																					
	1	**0.942**	0.614	**0.550**	0.459	**0.932**	0.673	1	**0.974**	0.195	**0.954**	0.163	**0.970**	0.191	1	**0.948**	0.062	**0.954**	0.060	**0.950**	0.060
	2	**0.994**	0.612	**0.856**	0.352	**0.998**	0.581	2	**0.996**	0.194	**0.990**	0.136	**0.984**	0.172	2	**0.966**	0.062	**0.968**	0.057	**0.960**	0.059
	3	**1.000**	0.613	**0.946**	0.309	**0.992**	0.525	3	**1.000**	0.195	**0.990**	0.129	**0.980**	0.166	3	**0.964**	0.062	**0.970**	0.057	**0.966**	0.059
	4	**1.000**	0.607	**0.982**	0.286	**0.980**	0.472	4	**0.996**	0.195	**0.994**	0.127	**0.984**	0.162	4	**0.946**	0.061	**0.956**	0.057	**0.950**	0.059
	5	**1.000**	0.611	**0.988**	0.271	**0.966**	0.427	5	**0.996**	0.195	**0.996**	0.126	**0.976**	0.164	5	**0.960**	0.062	**0.966**	0.057	**0.960**	0.059
	6	**1.000**	0.609	**0.994**	0.260	**0.964**	0.385	6	**0.998**	0.195	**0.992**	0.125	**0.970**	0.166	6	**0.956**	0.062	**0.958**	0.057	**0.956**	0.059
	7	**1.000**	0.609	**0.996**	0.254	**0.964**	0.347	7	**0.994**	0.195	**0.990**	0.126	**0.966**	0.169	7	**0.950**	0.062	**0.960**	0.057	**0.948**	0.059
	8	**1.000**	0.611	**0.994**	0.251	**0.974**	0.314	8	**0.996**	0.196	**0.994**	0.127	**0.968**	0.175	8	**0.962**	0.062	**0.974**	0.057	**0.966**	0.059
	9	**1.000**	0.619	**0.998**	0.247	**0.982**	0.288	9	**0.996**	0.195	**0.990**	0.127	**0.960**	0.176	9	**0.958**	0.062	**0.964**	0.057	**0.958**	0.059
	10	**1.000**	0.610	**1.000**	0.245	**0.988**	0.270	10	**0.990**	0.194	**0.992**	0.127	**0.970**	0.176	10	**0.964**	0.062	**0.966**	0.057	**0.962**	0.059
	11	**1.000**	0.609	**1.000**	0.246	**0.994**	0.263	11	**0.996**	0.195	**0.990**	0.127	**0.978**	0.173	11	**0.954**	0.062	**0.956**	0.057	**0.956**	0.059
Normal																					
	1	**0.960**	0.607	**0.960**	0.559	**0.960**	0.569	1	**0.942**	0.195	**0.942**	0.195	**0.942**	0.195	1	**0.942**	0.062	**0.942**	0.062	**0.942**	0.062
	2	**0.968**	0.612	**0.968**	0.500	**0.970**	0.532	2	**0.940**	0.194	**0.940**	0.193	**0.940**	0.193	2	**0.954**	0.062	**0.954**	0.062	**0.954**	0.062
	3	**0.972**	0.620	**0.966**	0.461	**0.962**	0.519	3	**0.956**	0.195	**0.958**	0.191	**0.958**	0.191	3	**0.962**	0.062	**0.962**	0.062	**0.962**	0.062
	4	**0.992**	0.611	**0.988**	0.431	**0.988**	0.507	4	**0.954**	0.196	**0.956**	0.187	**0.956**	0.188	4	**0.938**	0.061	**0.938**	0.061	**0.938**	0.061
	5	**0.994**	0.606	**0.986**	0.410	**0.980**	0.496	5	**0.948**	0.195	**0.954**	0.181	**0.954**	0.183	5	**0.954**	0.062	**0.954**	0.062	**0.954**	0.062
	6	**0.996**	0.611	**0.986**	0.392	**0.978**	0.490	6	**0.960**	0.195	**0.960**	0.177	**0.962**	0.182	6	**0.952**	0.062	**0.952**	0.062	**0.952**	0.062
	7	**0.998**	0.610	**0.986**	0.383	**0.986**	0.487	7	**0.964**	0.196	**0.976**	0.173	**0.972**	0.179	7	**0.942**	0.062	**0.942**	0.062	**0.942**	0.062
	8	**0.996**	0.611	**0.992**	0.376	**0.992**	0.477	8	**0.970**	0.195	**0.974**	0.174	**0.972**	0.181	8	**0.954**	0.062	**0.954**	0.061	**0.954**	0.061
	9	**0.998**	0.614	**0.996**	0.370	**0.986**	0.472	9	**0.964**	0.196	**0.972**	0.172	**0.970**	0.180	9	**0.952**	0.062	**0.952**	0.062	**0.952**	0.062
	10	**0.996**	0.613	**0.994**	0.364	**0.986**	0.464	10	**0.954**	0.194	**0.962**	0.171	**0.960**	0.179	10	**0.952**	0.062	**0.952**	0.062	**0.952**	0.062
	11	**1.000**	0.616	**0.994**	0.366	**0.990**	0.468	11	**0.974**	0.195	**0.976**	0.171	**0.976**	0.178	11	**0.948**	0.062	**0.948**	0.062	**0.948**	0.062

4.6 Example

To illustrate our proposed methodology, we use the data from 14 studies which assessed the effect of an anti-oxidant (acetylcysteine) in preventing contrast-induced nephropathy, a leading cause of acquired acute reduction in kidney function (Bagshaw and Ghali, 2004). The outcome of interest in each study was incidence of contrast-induced nephropathy, and so the parameter of interest was the odds ratio for the association between anti-oxidant usage and incidence of nephropathy. The summary data for each study is shown below.

Table 4.3: Summary results of 14 studies of acetylcysteine for prevention of contrast-induced nephropathy

Study	N	OR	CI
Allaqaband	85	1.23	(0.39, 3.89)
Baker	80	0.20	(0.04, 1.00)
Briguori	183	0.57	(0.20, 1.63)
Diaz-Sandova	54	0.11	(0.02, 0.54)
Durham	79	1.27	(0.45, 3.57)
Efrati	49	0.19	(0.01, 4.21)
Fung	91	1.37	(0.43, 4.32)
Goldenberg	80	1.30	(0.27, 6.21)
Kay	200	0.29	(0.09, 0.94)
Kefer	104	0.63	(0.10, 3.92)
MacNeill	43	0.11	(0.01, 0.97)
Oldemeyer	96	1.30	(0.28, 6.16)
Shyu	121	0.11	(0.02, 0.49)
Vallero	100	1.14	(0.27, 4.83)

A fixed effects analysis of this data results in a 95% confidence interval of (0.41, 0.87) for the (assumed) common odds ratio. However, significant heterogeneity was found in the study-level treatment effects (p=0.032). A random effects analysis, assuming that the logs of the study-level odds ratios are normally distributed, results in a somewhat wider confidence interval (0.32, 0.91). Below we show the resulting 95% confidence intervals for each of the 14 ordered study-level treatment effects. The three columns of confidence intervals correspond to the weighting methods discussed in this article, with the third column representing our

proposed procedure, which we have shown in simulations to have appropriate coverage, regardless of whether any or all of the true treatment effects are equal across studies. Even though we have some evidence to reject the fixed effects assumption, in this example it is particularly difficult, due to small sample sizes, to assess with any certainty whether or not any subsets of the study parameters are equal to one another, or whether the assumption of normal distribution for the log-odds-ratios is appropriate. We note that, in general, the intervals provided by Method 1 are essentially a re-ordering of the original study intervals, and thus do not provide substantially new information in terms of summarizing the treatment effects. The bootstrap intervals corresponding to Method 2, are noticeably narrower in some cases; however, it is concerning that the interval for $\theta^{(14)}$, (1.44, 9.56), excludes even the maximum estimated treatment effect (estimated odds ratio = 1.37 from the Fung study). Using our proposed weights (Method 3), we estimate that six of the fourteen studies exhibited significant treatment effects, while the remaining eight studies were found to be neutral. The confidence intervals for the 7th and 8th ordered treatment effects are (0.28, 1.01) and (0.30, 1.07), respectively. Using the conventional method of averaging the $(K/2)^{th}$ and $(K/2+1)^{th}$ ordered observations to estimate the median when K is an even number, we obtain a confidence interval of (0.29, 1.04) for the "median" treatment effect across these studies. This interval is wider than the previously reported random effects analysis, though our inference is free of any distributional assumptions regarding the true values of the study-level treatment effects. Furthermore, if the true distribution of the parameters is not symmetric on the log scale, then our estimate of the median treatment effect will not necessarily be directly comparable to the random effects analysis, which estimates the mean of the random-effects distribution.

In Figure 1, we present the 95% confidence intervals for each ordered element of $\{\Theta\}$, with point estimates given by the mean of the associated confidence distribution. For comparison, the confidence intervals for the fixed-effects and random-effects meta-analysis are denoted by the vertical solid and dashed lines, respectively. Our estimates for $\theta^{(7)}$ and $\theta^{(8)}$ are highlighted for comparison.

Table 4.4: 95% Confidence Intervals for ordered study-level treatment effects using nephropathy data

OS	CI (Method 1)	CI (Method 2)	CI (Method 3)
1	(0.02, 0.48)	(0.01, 0.13)	(0.01, 0.65)
2	(0.02, 0.51)	(0.03, 0.20)	(0.04, 0.71)
3	(0.01, 0.94)	(0.05, 0.28)	(0.08, 0.71)
4	(0.01, 4.64)	(0.07, 0.40)	(0.14, 0.74)
5	(0.04, 1.04)	(0.12, 0.54)	(0.18, 0.80)
6	(0.09, 0.94)	(0.17, 0.70)	(0.23, 0.88)
7	(0.19, 1.67)	(0.25, 0.91)	(0.28, 1.01)
8	(0.10, 3.85)	(0.33, 1.16)	(0.30, 1.07)
9	(0.27, 4.93)	(0.45, 1.46)	(0.31, 1.25)
10	(0.39, 3.94)	(0.56, 1.79)	(0.32, 1.44)
11	(0.45, 3.49)	(0.70, 2.27)	(0.31, 1.61)
12	(0.26, 6.06)	(0.87, 2.99)	(0.30, 1.87)
13	(0.28, 6.14)	(1.09, 4.38)	(0.30, 2.31)
14	(0.44, 4.26)	(1.44, 9.56)	(0.32, 4.07)

4.6.1 ECDF

While our proposed procedure was motivated by a desire to avoid making any assumptions about the existence or nature of the distribution of our quantity of interest $\{\Theta\}$, we note that a plot such as that given in Figure 1 may resemble an empirical CDF for the "true" distribution $F(\Theta)$. As sample size increases, the confidence distribution estimates for each $\theta^{(m)}$ converge to the true values $(\theta^{(1)}, \theta^{(2)}, ..., \theta^{(K)})$. If it can further be assumed that $(\theta^{(1)}, \theta^{(2)}, ..., \theta^{(K)})$ are a random sample from some overall distribution $F(\Theta)$, then it can be seen that $\tilde{\theta}^{(q)} = \theta^{(\lfloor qK \rfloor + 1)}$ will converge, as K grows large, to $F_{\Theta}^{-1}(q)$.

4.7 Discussion

In this paper, we introduce a unified framework which simultaneously addresses two important problems. By introducing a procedure for making inference on *any* ordered value of

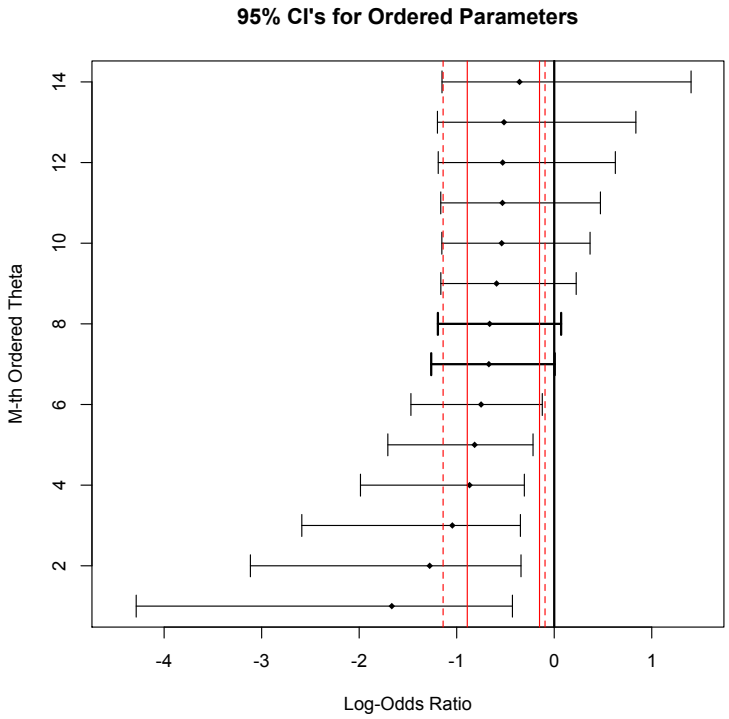


Figure 4.1: Confidence distribution estimates of treatment effects from 14 studies of acetylcysteine on nephropathy: Vertical and solid (dashed) lines represent 95% CI from fixed-effects (random-effects) meta-analysis

a set of parameters, we may provide a summary of the treatment effects observed over a collection of studies without having to rely on any assumptions about the nature of or relationship between those treatment effects, thus enabling a non-parametric, model-free form of meta-analysis. While the resulting confidence interval from such a procedure will likely be wider than those provided by methods with more restrictive assumptions, the general applicability of our new method is appealing and may serve as a good point of comparison, just as many analysts now present results corresponding to both fixed-effects and random-effects meta-analysis models. Alternatively, our procedure also allows us to make inference on the extreme values of a set of parameters, a well-established problem that has proven to be intractable with respect to many statistical approaches. By taking advantage of the flexibility afforded by confidence distributions as functional estimators, as well as a tuning technique that accounts for the unknown presence or absence of ties and near-ties in small-sample

settings, we are now able to provide valid inference in a wide variety of settings.

4.8 Appendix

Proof of Theorem 4.1. (i) The first two results follow immediately from (4.13) and the fact that $|\hat{\theta}_i - \theta^{(m)}| \leq |\hat{\theta}_i - \theta_i| + |\theta_i - \theta^{(m)}| = O_p(N^{-1/2})$ for any $\theta_i \in \Theta_{\mathcal{N}}$. We only need to prove (4.15).

Note that, $\hat{\theta}_i \sim (\theta_i, s_i^2)$, for any i, it follows that

$$\frac{\sum_{i\in\Theta_{\mathcal{N}}} c_i\hat{\theta}_i \big/ \sum_{i\in\Theta_{\mathcal{N}}} c_i - \sum_{i\in\Theta_{\mathcal{N}}} c_i\theta_i \big/ \sum_{i\in\Theta_{\mathcal{N}}} c_i}{\sqrt{\sum_{i\in\Theta_{\mathcal{N}}} c_i^2 s_i^2 \big/ \{\sum_{i\in\Theta_{\mathcal{N}}} c_i\}^2}} \sim N(0,1). \tag{4.21}$$

Again, from (4.13) and the fact that $|\theta_i - \theta^{(m)}| = O(N^{-1/2})$ for any $\theta_i \in \Theta_{\mathcal{N}}$, we have $\sum_{i=1}^K w_{i,(m)}\hat{\theta}_i = \sum_{i\in\Theta_{\mathcal{N}}} c_i\hat{\theta}_i + o_p(1)$, $\sum_{i=1}^K w_{i,(m)}^2 s_i^2 = \sum_{i\in\Theta_{\mathcal{N}}} c_i^2 s_i^2 + o_p(1)$, $\sum_{i=1}^K w_{i,(m)} = \sum_{i\in\Theta_{\mathcal{N}}} c_i + o_p(1)$ and $\sum_{i\in\Theta_{\mathcal{N}}} c_i\theta_i = \{\sum_{i\in\Theta_{\mathcal{N}}} c_i\}\,\theta^{(m)} + O(N^{-1/2})$. Thus, we have

$$\frac{\sum_{i=1}^K w_{i,(m)}\hat{\theta}_i \big/ \sum_{i=1}^K w_{i,(m)} - \theta^{(m)}}{\sqrt{\sum_{i=1}^K w_{i,(m)}^2 s_i^2 \big/ \{\sum_{i=1}^K w_{i,(m)}\}^2}} \to N(0,1), \quad \text{as } N \to \infty \tag{4.22}$$

On the other hand, since $\xi^* = \sum_{i=1}^K w_{i,(m)}\xi_i / \sum_{i=1}^K w_{i,(m)}$ and ξ_i are CD random variables from $N(\hat{\theta}_i, s_i^2)$, we have

$$\left.\frac{\xi^* - \sum_{i=1}^K w_{i,(m)}\hat{\theta}_i \big/ \sum_{i=1}^K w_{i,(m)}}{\sqrt{\sum_{i=1}^K w_{i,(m)}^2 s_i^2 \big/ \{\sum_{i=1}^K w_{i,(m)}\}^2}}\right| \hat{\Theta} \sim N(0,1). \tag{4.23}$$

It follows immediately the third result of (i).

(ii) Based on (4.22) and (4.23) and the definition of $H_*(t)$, we have, for any $0 < s < 1$

and as $N \to \infty$,

$$
\begin{aligned}
&P\left\{H_*(\theta^{(m)}) \le s\right\} \\
&\quad = P\left\{ P\left(\frac{\xi^* - \sum_{i=1}^K w_{i,(m)}\hat{\theta}_i / \sum_{i=1}^K w_{i,(m)}}{\sqrt{\sum_{i=1}^K w_{i,(m)}^2 s_i^2 / \{\sum_{i=1}^K w_{i,(m)}\}^2}} \le \frac{\theta^{(m)} - \sum_{i=1}^K w_{i,(m)}\hat{\theta}_i / \sum_{i=1}^K w_{i,(m)}}{\sqrt{\sum_{i=1}^K w_{i,(m)}^2 s_i^2 / \{\sum_{i=1}^K w_{i,(m)}\}^2}} \middle| \hat{\Theta} \right) \le s \right\} \\
&\quad = P\left\{ \frac{\theta^{(m)} - \sum_{i=1}^K w_{i,(m)}\hat{\theta}_i / \sum_{i=1}^K w_{i,(m)}}{\sqrt{\sum_{i=1}^K w_{i,(m)}^2 s_i^2 / \{\sum_{i=1}^K w_{i,(m)}\}^2}} \le \Phi^{-1}(s) \right\} \to \Phi(\Phi^{-1}(s)) = s. \qquad (4.24)
\end{aligned}
$$

Thus, $H_*(\theta^{(m)}) \to Unif[0,1]$, as $N \to \infty$. The conclusion of (ii) follows.

Proof of Lemma 4.1 Recall that the condition described in (4.13) is as follows:

$$\lim_{n\to\infty} w_{i,(m)} = \begin{cases} c_i & \text{if } \imath \in \Theta_T^{(m)}, \\ 0 & \text{if } \imath \notin \Theta_T^{(m)}, \end{cases} \quad \text{for } i = 1, 2, \ldots, K \ .$$

Without loss of generality, let $\theta_1 < \theta_2 < \ldots < \theta_K$. Also let $\hat{\theta}_j \sim N(\theta_j, \sigma_j^2/n_j)$ for each j and define $\hat{\theta}^{(j)} : \hat{\theta}^{(1)} \le \hat{\theta}^{(2)} \le ... \le \hat{\theta}^{(K)}$. Furthermore, suppose we are interested in θ_m.

Recall $w_{m,(m)}^{[1]} = \mathbf{1}\{\hat{\theta}_m = \hat{\theta}^{(m)}\}$ is a binary random variable that equals 1 with probability $P\{\hat{\theta}_i = \hat{\theta}^{(m)}\}$.

$$
\begin{aligned}
P\{\hat{\theta}_m = \hat{\theta}^{(m)}\} &< \prod_{i<m} [P\{\hat{\theta}_i < \hat{\theta}_m\}] \prod_{j>m} [P\{\hat{\theta}_j > \hat{\theta}_m\}] \\
&= \int \prod_{i<m} [P\{\hat{\theta}_i < c\}] P\{\hat{\theta}_m = c\} \prod_{j>m} [P\{\hat{\theta}_j > c\}] \, \mathrm{d}c \\
&= \int \prod_{i<m} [\Phi(\frac{c-\theta_i}{\sigma_i/\sqrt{n_i}})] \phi(\frac{c-\theta_m}{\sigma_m/\sqrt{n_m}}) \prod_{j>m} [\Phi(\frac{\theta_j - c}{\sigma_j/\sqrt{n_j}})] \, \mathrm{d}c \qquad (4.25) \\
&< \int_{\theta_{m-1}+\epsilon}^{\theta_{m+1}-\epsilon} \prod_{i<m} [\Phi(\frac{c-\theta_i}{\sigma_i/\sqrt{n_i}})] \phi(\frac{c-\theta_m}{\sigma_m/\sqrt{n_m}}) \prod_{j>m} [\Phi(\frac{\theta_j - c}{\sigma_j/\sqrt{n_j}})] \, \mathrm{d}c \\
&\to \int_{\theta_{m-1}+\epsilon}^{\theta_{m+1}-\epsilon} \phi(\frac{c-\theta_m}{\sigma_m/\sqrt{n_m}}) \, \mathrm{d}c \to 1
\end{aligned}
$$

Thus $w_{m,(m)}^{[1]}$ converges in probability to 1.

Because we have that $w^{[1]}_{m,(m)} \to 1$ and $\sum_i w^{[1]}_{i,(m)} = 1$, then $w^{[1]}_{i,(m)} \to 0 \forall i \neq m$, thus satisfying (4.13).

Noting that $\hat{\theta}_j \sim N(\theta_j, \sigma_j^2/n_j)$ and, unconditionally, $\xi_j \sim N(\theta_j, 2\sigma_j^2/n_j)$, we can replace each σ_j^2 with $2\sigma_j^2$ in the proof above, and the result remains unchanged.

Recall that $w^{[3]}_{i,(m)} = \mathbf{1}\{-b_L \leq (\xi_i - \xi^{(m)}) \leq b_R\}$, where $(b_L, b_R) \propto \tau_N, \tau_N = O(N^{-\delta}), \delta \in (0, \frac{1}{2})$. For $i = m$, we use the argument above that $P\{\xi_m = \xi^{(m)}\} \to 1$, and so $\mathcal{K}\left(\frac{\xi_i - \xi^{(m)}}{\tau_N}\right) \to \mathcal{K}\left(\frac{0}{\tau_N}\right) = 1$. For $i \neq m$, $(\xi_i - \xi^{(m)})$ converges in probability to $D_i = \theta_i - \theta_m$. For $i < m$, $D_i/\tau_N \to -\infty$, and thus $\mathcal{K}\left(\frac{\xi_i - \xi^{(m)}}{\tau_N}\right) \to 0$. Similarly, for $i > m$, $D_i/\tau_N \to +\infty$, and thus $\mathcal{K}\left(\frac{\xi_i - \xi^{(m)}}{\tau_N}\right) \to 0$. Thus, we have satisfied (4.13).

Proof of Lemma 4.2. Recall that $w^{[3]}_{i,(m)} = \mathbf{1}\{-b_L \leq (\xi_i - \xi^{(m)}) \leq b_R\}$, where $(b_L, b_R) \propto \tau_N$, and $\tau_N = O(N^{-\delta})$, with $0 < \delta < 1/2$. Let us denote $c_L = b_L/\tau_N, c_R = b_R/\tau_N$, so that $c_L, c_R = O(1)$.

Now $w^{[3]}_{i,(m)} = \mathbf{1}\{-b_L \leq (\xi_i - \xi^{(m)}) \leq b_R\} = \mathbf{1}\{-c_L \leq \frac{(\xi_i - \xi^{(m)})}{\tau_N} \leq c_R\}$. Note that in general, $\xi_i = \hat{\theta}_i + \tilde{\epsilon}_n$, and $\hat{\theta}_i = \theta_i + \epsilon_n$, where both $\tilde{\epsilon}_n$ and $\epsilon_n = O(N^{-1/2})$. Substituting, we have that $(\xi_i - \xi^{(m)}) = \theta_i - \theta^{(m)} + \tilde{\epsilon}_{in} - \tilde{\epsilon}^{(m)}_n + \epsilon_{in} - \epsilon^{(m)}_n = \theta_i - \theta^{(m)} + O(N^{-1/2})$.

Thus, when $i \in \Theta^{(m)}_T, \theta_i = \theta^{(m)}$, then $(\xi_i - \xi^{(m)}) = O(N^{-1/2}), \frac{(\xi_i - \xi^{(m)})}{\tau_N} = O(N^{-1/2+\delta}) = o(1) \Rightarrow \frac{(\xi_i - \xi^{(m)})}{\tau_N} \to 0$ as $N \to \infty$, and so $w^{[3]}_{i,(m)} \to 1$ as $N \to \infty$.

First, we address the conventional case where θ_i are constant in N.

When $i \notin \Theta^{(m)}_T, \theta_i \neq \theta^{(m)}$, then $(\xi_i - \xi^{(m)}) = \theta_i - \theta^{(m)} + O(N^{-1/2}) = O(1), \frac{(\xi_i - \xi^{(m)})}{\tau_N} = O(N^{\delta}) \Rightarrow \frac{(\xi_i - \xi^{(m)})}{\tau_N} \to \infty$ as $N \to \infty$, and so $w^{[3]}_{i,(m)} \to 0$ as $N \to \infty$.

Now, if we instead have that $d_m = \min_{j \notin \Theta_T^{(m)}} |\theta_j - \theta^{(m)}|$ is $O(N^{-\delta^*})$, for any $\delta^* \in (0, 1/2)$, our only requirement is that the convergence rate for τ_N, δ must be restricted to $(\delta^*, 1/2)$.

In this case, when $i \notin \Theta_T^{(m)}, \theta_i \neq \theta^{(m)}$, then $(\xi_i - \xi^{(m)}) = O(N^{-\delta^*})$, $\frac{(\xi_i - \xi^{(m)})}{\tau_N} = O(N^{\delta - \delta^*}) \Rightarrow \frac{(\xi_i - \xi^{(m)})}{\tau_N} \to \infty$ as $N \to \infty$, and so $w_{i,(m)}^{[3]} \to 0$ as $N \to \infty$.

References

Agresti, A. (1990). *Categorical data analysis*. Wiley series in probability and mathematical statistics. Applied probability and statistics. Wiley.

Balasubramanian, R. and Lagakos, S. (2010). Estimating hiv incidence based on combined prevalence testing. *Biometrics* **66,** 1–10.

Beta-Blocker Evaluation of Survival Trial Investigators (2001). A trial of the beta-blocker bucindolol in patients with advanced chronic heart failure. *New England Journal of Medicine* **344,** 1659–1667.

Bickel, P. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. *The Annals of Statistics* pages 1071–1095.

Boers, M., Brooks, P., Fries, J. F., Simon, L. S., Strand, V., and Tugwell, P. (2010). A first step to assess harm and benefit in clinical trials in one scale. *Journal of clinical epidemiology* **63,** 627–632.

Bonetti, M. and Gelber, R. (2004). Patterns of treatment effects in subsets of patients in clinical trials. *Biostatistics* **5,** 465–481.

Bonetti, M., Gelber, R., et al. (2000). A graphical method to assess treatment-covariate interactions using the cox model on subsets of the data. *Statistics in medicine* **19,** 2595–2609.

Brookmeyer, R. (2009). Should biomarker estimates of hiv incidence be adjusted? *Aids* **23,** 485–491.

Brookmeyer, R. and Quinn, T. (1995). Estimation of current human immunodeficiency virus incidence rates from a cross-sectional survey using early diagnostic tests. *American journal of epidemiology* **141,** 166–172.

Cai, T., Tian, L., Uno, H., Solomon, S., and Wei, L. (2010). Calibrating parametric subject-specific risk estimation. *Biometrika* **97,** 389–404.

Cai, T., Tian, L., and Wei, L. (2005). Semiparametric box–cox power transformation models for censored survival observations. *Biometrika* **92,** 619–632.

Cai, T., Tian, L., Wong, P., and Wei, L. (2011). Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics* **12,** 270–282.

Cheng, S. C., Wei, L. J., and Ying, Z. (1995). Analysis of transformation models with censored data. *Biometrika* **82,** 835–845.

Chuang-Stein, C., Mohberg, N. R., and Sinkula, M. S. (1991). Three measures for simultaneously evaluating benefits and risks using categorical data from clinical trials. *Statistics in Medicine* **10,** 1349–1359.

Cox, D. (1958). Some problems with statistical inference. *The Annals of Mathematical Statistics* **29,** 357–372.

Cox, D. (2006). Some problems with statistical inference. *London: Cambridge University Press* .

Domanski, M., Krause-Steinrauf, H., Massie, B., Deedwania, P., Follmann, D., Kovar, D., Murray, D., Oren, R., Rosenberg, Y., Young, J., et al. (2003). A comparative analysis of the results from 4 trials of [beta]-blocker therapy for heart failure: Best, cibis-ii, merit-hf, and copernicus. *Journal of cardiac failure* **9,** 354–363.

Edwardes, M. and Baltzan, M. (2000). The generalization of the odds ratio, risk ratio and risk difference to r x k tables. *Statistics in medicine* **19,** 1901–1914.

Edwardes, M. D. d. (1995). A confidence interval for pr(x < y) - pr(x > y) estimated from simple cluster samples. *Biometrics* **51,** pp. 571–578.

Efron, B. (1993). Bayes and likelihood calculations from confidence intervals. *Biometrika* **80,** 3–26.

Efron, B. (1998). R.a.fisher in the 21st century. *Statistical Science* **13,** 95–122.

Ferreira-González, I., Permanyer-Miralda, G., Domingo-Salvany, A., Busse, J. W., Heels-Ansdell, D., and Montori, V. M. (2007). Problems with use of composite end points in cardiovascular trials: systematic review of randomised controlled trials. *BMJ* **334,** 786.

Fine, J. P. and Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association* **94,** 496–509.

Fisher, R. (1930). Inverse probability. *Proc. Cambridge Pilos. Soc.* **26,** 528–535.

Fisher, R. (1973). Statistical methods and scientific inference. *New York: Hafner Press* .

Follmann, D. A. (2002). Regression analysis based on pairwise ordering of patients' clinical histories. *Statistics in Medicine* **21,** 3353–3367.

Friedman, L., Furberg, C., and DeMets, D. (2010). *Fundamentals of Clinical Trials*. Springer.

Gelber, R. D., Cole, B. F., Gelber, S., and Goldhirsch, A. (1995). Comparing treatments using quality-adjusted survival: The q-twist method. *The American Statistician* **49,** 161–169.

Gelber, R. D., Goldhirsch, A., Cole, B. F., Wieand, H. S., Schroeder, G., and Krook, J. E. (1996). A quality-adjusted time without symptoms or toxicity (q-twist) analysis of adjuvant radiation therapy and chemotherapy for resectable rectal cancer. *Journal of the National Cancer Institute* **88,** 1039–1045.

Hall, P. and Miller, H. (2010). Bootstrap confidence intervals and hypothesis tests for extrema of parameters. *Biometrika* **97,** 881–892.

Hargrove, J., Humphrey, J., Mutasa, K., Parekh, B., McDougal, J., Ntozini, R., Chidawanyika, H., Moulton, L., Ward, B., Nathoo, K., et al. (2008). Improved hiv-1 incidence estimates using the bed capture enzyme immunoassay. *Aids* **22,** 511.

Huang, Y., Sullivan Pepe, M., and Feng, Z. (2007). Evaluating the predictiveness of a continuous marker. *Biometrics* **63,** 1181–1188.

Janes, H., Pepe, M., Bossuyt, P., and Barlow, W. (2011). Measuring the performance of markers for guiding treatment decisions. *Annals of internal medicine* **154,** 253.

Janssen, R., Satten, G., Stramer, S., Rawal, B., O'Brien, T., Weiblen, B., Hecht, F., Jack, N., Cleghorn, F., Kahn, J., et al. (1998). New testing strategy to detect early hiv-1 infection for use in incidence estimates and for clinical and prevention purposes. *JAMA: the journal of the American Medical Association* **280,** 42–48.

Kent, D. and Hayward, R. (2007). Limitations of applying summary results of clinical trials to individual patients. *JAMA: the journal of the American Medical Association* **298,** 1209–1212.

Li, Y., Tian, L., and Wei, L. (2011). Estimating subject-specific dependent competing risk profile with censored event time observations. *Biometrics* **67,** 427–435.

Lin, D., Wei, L., and Ying, Z. (1993). Checking the cox model with cumulative sums of martingale-based residuals. *Biometrika* **80,** 557–572.

Lin, D. Y. and Wei, L. J. (1989). The robust inference for the cox proportional hazards model. *Journal of the American Statistical Association* **84,** pp. 1074–1078.

Lui, K.-J. (2002). Notes on estimation of the general odds ratio and the general risk difference for paired-sample data. *Biometrical Journal* **44,** 957–968.

Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *The Annals of Statistics* **21,** 255–285.

Marinda, E., Hargrove, J., Preiser, W., Slabbert, H., van Zyl, G., Levin, J., Moulton, L., Welte, A., and Humphrey, J. (2010). Significantly diminished long-term specificity of the bed capture enzyme immunoassay among patients with hiv-1 with very low cd4 counts and those on antiretroviral therapy. *JAIDS Journal of Acquired Immune Deficiency Syndromes* **53,** 496.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological)* **42,** pp. 109–142.

Mcdougal, J., Parekh, B., Peterson, M., Branson, B., Dobbs, T., Ackers, M., and Gurwith, M. (2006). Comparison of hiv type 1 incidence observed during longitudinal follow-up with incidence estimated by cross-sectional analysis using the bed capture enzyme immunoassay. *AIDS Research & Human Retroviruses* **22,** 945–952.

Parekh, B., Kennedy, M., Dobbs, T., Pau, C., Byers, R., Green, T., Hu, D., Vanichseni, S., Young, N., Choopanya, K., et al. (2002). Quantitative detection of increasing hiv type 1 antibodies after seroconversion: a simple assay for detecting recent hiv infection and estimating incidence. *AIDS research and human retroviruses* **18,** 295–307.

Park, Y. and Wei, L. (2003a). Estimating subject-specific survival functions under the accelerated failure time model. *Biometrika* **90,** 717–723.

Park, Y. and Wei, L. J. (2003b). Estimating subjectspecific survival functions under the accelerated failure time model. *Biometrika* **90,** 717–723.

Pepe, M. (2004). *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press, USA.

Pepe, M., Feng, Z., Huang, Y., Longton, G., Prentice, R., Thompson, I., and Zheng, Y. (2008). Integrating the predictiveness of a marker with its performance as a classifier. *American journal of epidemiology* **167,** 362–368.

Peterson, P., Rumsfeld, J., Liang, L., Hernandez, A., Peterson, E., Fonarow, G., Masoudi, F., et al. (2010). Treatment and risk in heart failure. *Circulation: Cardiovascular Quality and Outcomes* **3,** 309–315.

Sakarovitch, C., Rouet, F., Murphy, G., Minga, A., Alioum, A., Dabis, F., Costagliola, D., Salamon, R., Parry, J., and Barin, F. (2007). Do tests devised to detect recent hiv-1 infection provide reliable estimates of incidence in africa? *JAIDS Journal of Acquired Immune Deficiency Syndromes* **45,** 115.

Schweder, T. and Hjort, N. (2002). Confidence and likelihood. *Scandinavian Journal of Statistics* **29,** 309–332.

Signorovitch, J. (2007). *Identifying informative biological markers in high-dimensional genomic data and clinical trials*. PhD thesis, Harvard University.

Simonoff, J. S., Hochberg, Y., and Reiser, B. (1986). Alternative estimation procedures for pr(x < y) in categorized data. *Biometrics* **42,** pp. 895–907.

Singh, K., Xie, M., and Strawderman, W. (2005). Combining information from independent sources through confidence distributions. *The Annals of Statistics* **33,** 159–183.

Singh, K., Xie, M., and Strawderman, W. (2007). Confidence distribution (cd): Distribution estimator of a parameter. *Lecture Notes-Monograph Series* pages 132–150.

Song, X. and Pepe, M. (2004). Evaluating markers for selecting a patient's treatment. *Biometrics* **60,** 874–883.

Sutton, A. J. and Higgins, J. P. T. (2008). Recent developments in meta-analysis. *Statistics in Medicine* **27,** 625–650.

Therneau, T. and Grambsch, P. (2000). *Modeling Survival Data: Extending the Cox Model*. Statistics for Biology and Health. Springer.

Tomlinson, G. and Detsky, A. S. (2010). Composite end points in randomized trials. *JAMA: The Journal of the American Medical Association* **303,** 267–268.

Uno, H., Cai, T., Tian, L., and Wei, L. J. (2007a). Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association* **102,** 527–537.

Uno, H., Cai, T., Tian, L., and Wei, L. J. (2007b). Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association* **102,** 527–537.

Wang, R. and Lagakos, S. (2009). On the use of adjusted cross-sectional estimators of hiv incidence. *JAIDS Journal of Acquired Immune Deficiency Syndromes* **52,** 538.

Wang, R. and Lagakos, S. (2010). Augmented cross-sectional prevalence testing for estimating hiv incidence. *Biometrics* **66,** 864–874.

Wang, R., Lagakos, S., Ware, J., Hunter, D., and Drazen, J. (2007). Statistics in medicinereporting of subgroup analyses in clinical trials. *New England Journal of Medicine* **357,** 2189–2194.

Wei, L. J. and Johnson, W. E. (1985). Combining dependent tests with incomplete repeated measurements. *Biometrika* **72,** 359–364.

Wei, L. J., Lin, D. Y., and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association* **84,** pp. 1065–1073.

Welte, A., McWalter, T., and Bärnighausen, T. (2009). A simplified formula for inferring hiv incidence from cross-sectional surveys using a test for recent infection. *AIDS research and human retroviruses* **25,** 125–126.

Woolf, B. et al. (1955). On estimating the relation between blood group and disease. *Ann Hum Genet* **19,** 251–253.

Wu, C. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics* **14,** 1261–1295.

Xie, M. and Singh, K. (2012). Confidence distribution, the frequentist distribution estimator of a parameter — a Review. *Int. Statist. Rev.* page In press (Invited review article with discussions).

Xie, M., Singh, K., and Zhang, C.-H. (2009). Confidence intervals for population ranks in the presence of ties and near ties. *Journal of the American Statistical Association* **104,** 775–788.

Zheng, Y., Cai, T., and Feng, Z. (2006). Application of the time-dependent roc curves for prognostic accuracy with multiple biomarkers. *Biometrics* **62,** 279–287.